# REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-03-

0481



| 1. AGENCY USE ONLY (Leave blank) | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | | FINAL REPORT 1 JULY 99 - 30 JUN 02 |

**4. TITLE AND SUBTITLE**
Mental Representation of Auditory Sources

**5. FUNDING NUMBERS**
F496209910293

**6. AUTHOR(S)**
Stephen Lakatos
Gary Scavone
James Beauchamp

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Washington State University Vancouver
14204 NE Salmon Creek Avenue
Vancouver, WA 98686

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

**11. SUPPLEMENTARY NOTES**

**12a. DISTRIBUTION / AVAILABILITY STATEMENT**

Approve for Public Release: Distribution Unlimited

20040105 076

**13. ABSTRACT** ***(Maximum 200 Words)***

The human auditory system possesses a remarkable ability to differentiate acoustic signals by the vibrational characteristics of their underlying sound sources. Understanding how listeners can detect, discriminate, classify, and remember acoustic source properties formed this project's overall goal. Using methods of signal detection, preliminary studies determined how listeners' sensitivity to auditory signals depends on whether attention is first directed to their acoustic features. Additional studies used perceptual mapping, new spectral measures, and novel data collection techniques to determine the acoustic cues listeners use to judge auditory perceptual similarity. A fundamental problem in auditory perception is to understand how listeners can perceive a sound source to be constant across wide variations in the sounds that the source can produce. Studies using simple and complex resonators demonstrated that listeners can represent the invariant properties of sound sources despite considerable variability in their excitation characteristics. Our ability to recognize previously heard sounds indicates that we encode features of acoustic sources in memory. Additional experiments used recognition and recall tasks, as well as measures of auditory "realism," to determine what cues persist in working and long-term memory. In sum, our research has shed important initial light on the human representation of auditory source properties.

| 14. SUBJECT TERMS | | | 15. NUMBER OF PAGES |
|---|---|---|---|
| | | | 16. PRICE CODE |
| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |

# FINAL PERFORMANCE REPORT

**Stephen Lakatos, Ph.D. (Principal Investigator)**
**Washington State University**
**14204 NE Salmon Creek Avenue**
**Vancouver, WA 98686**

**Gary Scavone, Ph.D.**
**Center for Computer Research in Music and Acoustics**
**The Knoll**
**Stanford University**
**Stanford, CA 94305**

**James Beauchamp, Ph.D.**
**Department of Electrical Engineering**
**University of Illinois at Urbana-Champaign**
**Urbana, IL 61801**

**AFOSR Agreement Number: F496209910293**

## 2. Objectives (changes to original objectives only)

No changes since first progress report.

## 3. Status of Effort

The human auditory system possesses a remarkable ability to differentiate acoustic signals by the vibrational characteristics of their underlying sound sources. Understanding how listeners can detect, discriminate, classify, and remember acoustic source properties formed this project's overall goal. Using methods of signal detection, preliminary studies determined how listeners' sensitivity to auditory signals depends on whether attention is first directed to their acoustic features. Additional studies used perceptual mapping, new spectral measures, and novel data collection techniques to determine the acoustic cues listeners use to judge auditory perceptual similarity. A fundamental problem in auditory perception is to understand how listeners can perceive a sound source to be constant across wide variations in the sounds that the source can produce. Studies using simple and complex resonators demonstrated that listeners can represent the invariant properties of sound sources despite considerable variability in their excitation characteristics. Our ability to recognize previously heard sounds indicates that we encode features of acoustic sources in memory. Additional experiments used recognition and recall tasks, as well as measures of auditory "realism," to determine what cues persist in working and long-term memory. In sum, our research has shed important initial light on the human representation of auditory source properties.

# 4. Accomplishments/New Findings

## 4.1 Selective Attention to Sound Source Properties

**SUMMARY:** Two experiments tested listeners' ability to attend selectively to the properties of a physical model comprising collisions between multiple independent sound-producing objects. A probe signal paradigm measured attention to two properties – resonant frequency and number of colliding objects. Listeners completed a baseline task measuring absolute sensitivity at each stimulus against a background noise. Subsequently, stimuli served as both cues and targets; cue validity was probabilistic. When cue and target were generated by the same object (Experiment 1), greater detectability occurred with valid cues for both resonant frequency and object number, implying the presence of attentional mechanisms for these properties. When cue and target were generated by different objects (Experiment 2), selective attention persisted for object number but not for resonant frequency.

### 4.1.1 Background

There is growing evidence that the auditory system can parse and represent the physical properties of sound-generating sources. For example, Warren and Verbrugge (1984) discovered that listeners can distinguish between "breaking" and "bouncing" sounds of glass bottles by attending to their higher-order temporal properties. Lakatos, McAdams, and Caussé (1997) used a crossmodal-matching task to measure listeners' ability to detect differences in the width-height ratios of steel and wooden bars and found that listeners are able to distinguish such dimensions by parsing the vibrational modes of the bars.

If physical properties of sound sources can be presented auditorially, can we attend selectively to them? This is a difficult question to address with natural sources, since it is often cumbersome to isolate individual physical attributes. Recently developed tools that can permit such manipulations, however, are physical models of sound sources implemented digitally. A number of such models is now available, ranging from exhaustive syntheses based on finite difference solutions of sound sources to computationally efficient, heuristics-based techniques that model the most perceptually salient acoustic properties of a source. To measure selective attention to source properties, we used an algorithm designed by Cook (1997) that models random collisions between particles using parametric stochastic synthesis.

Studies of auditory attention using variants of the auditory probe signal paradigm (Greenberg and Larkin, 1968) have consistently demonstrated selective attention to frequency, spatial location, and intensity (see Scharf, 1998). Recent neuropsychological findings support the notion of sensory gating mechanisms involved in attending to these attributes (e.g., Alcaini *et al.*, 1994). In the current work, we used the probe signal paradigm to examine selective attention not to frequency or spatial location, but to the acoustic properties of a physical model of colliding objects. Using a two-interval forced choice method, we first determined signal levels that produced a uniform detectability across all stimuli for individual subjects. We then measured detectability of the same stimuli under conditions in which one signal was expected (attended) and the other was not (unattended). Attention was manipulated by cueing subjects before each trial with a weak suprathreshold signal of the kind that was likely to be presented. Greater detectability for attended signals would indicate an effect of selective attention for acoustic properties of the cue.

This is a 2nd p.4, identical to p.3

# 4. Accomplishments/New Findings

## 4.1 Selective Attention to Sound Source Properties

**SUMMARY:** Two experiments tested listeners' ability to attend selectively to the properties of a physical model comprising collisions between multiple independent sound-producing objects. A probe signal paradigm measured attention to two properties – resonant frequency and number of colliding objects. Listeners completed a baseline task measuring absolute sensitivity at each stimulus against a background noise. Subsequently, stimuli served as both cues and targets; cue validity was probabilistic. When cue and target were generated by the same object (Experiment 1), greater detectability occurred with valid cues for both resonant frequency and object number, implying the presence of attentional mechanisms for these properties. When cue and target were generated by different objects (Experiment 2), selective attention persisted for object number but not for resonant frequency.

### 4.1.1 Background

There is growing evidence that the auditory system can parse and represent the physical properties of sound-generating sources. For example, Warren and Verbrugge (1984) discovered that listeners can distinguish between "breaking" and "bouncing" sounds of glass bottles by attending to their higher-order temporal properties. Lakatos, McAdams, and Caussé (1997) used a crossmodal-matching task to measure listeners' ability to detect differences in the width-height ratios of steel and wooden bars and found that listeners are able to distinguish such dimensions by parsing the vibrational modes of the bars.

If physical properties of sound sources can be presented auditorially, can we attend selectively to them? This is a difficult question to address with natural sources, since it is often cumbersome to isolate individual physical attributes. Recently developed tools that can permit such manipulations, however, are physical models of sound sources implemented digitally. A number of such models is now available, ranging from exhaustive syntheses based on finite difference solutions of sound sources to computationally efficient, heuristics-based techniques that model the most perceptually salient acoustic properties of a source. To measure selective attention to source properties, we used an algorithm designed by Cook (1997) that models random collisions between particles using parametric stochastic synthesis.

Studies of auditory attention using variants of the auditory probe signal paradigm (Greenberg and Larkin, 1968) have consistently demonstrated selective attention to frequency, spatial location, and intensity (see Scharf, 1998). Recent neuropsychological findings support the notion of sensory gating mechanisms involved in attending to these attributes (e.g., Alcaini *et al.*, 1994). In the current work, we used the probe signal paradigm to examine selective attention not to frequency or spatial location, but to the acoustic properties of a physical model of colliding objects. Using a two-interval forced choice method, we first determined signal levels that produced a uniform detectability across all stimuli for individual subjects. We then measured detectability of the same stimuli under conditions in which one signal was expected (attended) and the other was not (unattended). Attention was manipulated by cueing subjects before each trial with a weak suprathreshold signal of the kind that was likely to be presented. Greater detectability for attended signals would indicate an effect of selective attention for acoustic properties of the cue.

Same as p. 5

### 4.1.2 Physically Informed Sonic Modeling (PhISM)

The goal of physically informed sonic modeling (PhISM) is to couple physical simulations to efficient synthesis techniques (Cook, 1997). PhISEM (physically informed stochastic event modeling) is a PhISM algorithm based on particle models. In PhISEM, models of particles in containers were solved numerically using the basic Newtonian equations of motion. Simulations with varying numbers of particles and damping (loss of energy when particles collide with each other or the container) were run and statistics were collected about the likelihood of a sound-producing collision, the overall decay in sound energy, etc. Sound is produced only by particles hitting the container shell, because collisions between particles do not couple efficiently to the radiated sound.

The resulting PhISEM synthesis algorithm reduces the behavior of particle systems to a statistical process in which parameters relate directly to the parameters collected in the direct simulations. System energy, which represents the total kinetic energy in the system, decays exponentially. This exponential decay is rapid for systems with high damping. There is a Poisson probability of sound-producing collisions with a high waiting time for few objects and a low waiting time for many objects. This model approaches the ideal for larger numbers of particles. Sound-producing events are modeled as a short exponentially decaying of white noise, and the system resonances are modeled using biquad resonance filters. Even though the original models studied were particles within a sphere (beans within the gourd of a virtual maraca), PhISEM extends well to other systems with multiple independent sound-producing objects. From a psychoacoustic standpoint, the PhISEM model is appealing because it permits control over several physical parameters including the number of colliding objects, the damping properties, or the resonant frequency of the gourd (or chimes) themselves. We selected two parameters - the number of objects and the resonant frequency - to determine whether listeners could attend to them.

The bamboo wind chime and guiro models are different in two important ways. In the guiro, the resonance is fixed, modeling the gourd, whereas in the wind chimes, the resonance parameter is an average center resonance of a distribution of chime frequencies. Each time there is a collision in the chimes, a new frequency is allocated randomly +/- 20 percent around the target center resonance. The number of objects parameter expresses the (statistical) number of bamboo cylinders in the wind chimes. The guiro number of objects parameter applies to how many serrations are caught by the stick as the guiro is scraped.

### 4.1.3 Experiment 1

Subjects were ten undergraduates between the ages of 18 and 40 recruited from Washington State University and compensated for their efforts. None reported any hearing problems.

Four stimuli were generated with the PhISEM model of the bamboo chimes by crossing two levels of resonant frequency - "low" (1.6, 1.7 kHz spectral centroid) and "high" (4.2 kHz centroid) with two levels of object number - "few" (4-6 objects) and "many" (28-32 objects). We selected these values to generate sounds that were clearly able to be discriminated by listeners. A constant shake energy was used to generate all four stimuli. Figures 1a, b, c, and d plot the spectra for the four bamboo chime sounds averaged across the full duration of the sounds. Spectral centroid values are calculated with a formula

### 4.1.2 Physically Informed Sonic Modeling (PhISM)

The goal of physically informed sonic modeling (PhISM) is to couple physical simulations to efficient synthesis techniques (Cook, 1997). PhISEM (physically informed stochastic event modeling) is a PhISM algorithm based on particle models. In PhISEM, models of particles in containers were solved numerically using the basic Newtonian equations of motion. Simulations with varying numbers of particles and damping (loss of energy when particles collide with each other or the container) were run and statistics were collected about the likelihood of a sound-producing collision, the overall decay in sound energy, etc. Sound is produced only by particles hitting the container shell, because collisions between particles do not couple efficiently to the radiated sound.

The resulting PhISEM synthesis algorithm reduces the behavior of particle systems to a statistical process in which parameters relate directly to the parameters collected in the direct simulations. System energy, which represents the total kinetic energy in the system, decays exponentially. This exponential decay is rapid for systems with high damping. There is a Poisson probability of sound-producing collisions with a high waiting time for few objects and a low waiting time for many objects. This model approaches the ideal for larger numbers of particles. Sound-producing events are modeled as a short exponentially decaying of white noise, and the system resonances are modeled using biquad resonance filters. Even though the original models studied were particles within a sphere (beans within the gourd of a virtual maraca), PhISEM extends well to other systems with multiple independent sound-producing objects. From a psychoacoustic standpoint, the PhISEM model is appealing because it permits control over several physical parameters including the number of colliding objects, the damping properties, or the resonant frequency of the gourd (or chimes) themselves. We selected two parameters - the number of objects and the resonant frequency - to determine whether listeners could attend to them.

The bamboo wind chime and guiro models are different in two important ways. In the guiro, the resonance is fixed, modeling the gourd, whereas in the wind chimes, the resonance parameter is an average center resonance of a distribution of chime frequencies. Each time there is a collision in the chimes, a new frequency is allocated randomly +/- 20 percent around the target center resonance. The number of objects parameter expresses the (statistical) number of bamboo cylinders in the wind chimes. The guiro number of objects parameter applies to how many serrations are caught by the stick as the guiro is scraped.

### 4.1.3 Experiment 1

Subjects were ten undergraduates between the ages of 18 and 40 recruited from Washington State University and compensated for their efforts. None reported any hearing problems.

Four stimuli were generated with the PhISEM model of the bamboo chimes by crossing two levels of resonant frequency - "low" (1.6, 1.7 kHz spectral centroid) and "high" (4.2 kHz centroid) with two levels of object number - "few" (4-6 objects) and "many" (28-32 objects). We selected these values to generate sounds that were clearly able to be discriminated by listeners. A constant shake energy was used to generate all four stimuli. Figures 1a, b, c, and d plot the spectra for the four bamboo chime sounds averaged across the full duration of the sounds. Spectral centroid values are calculated with a formula

used by Beauchamp (1995). There was little spectral overlap between sounds from sources with low and high resonant frequencies. Those with similar frequencies share several resonance modes, although their peaks shift somewhat depending on the number of colliding objects.

The probe signal paradigm had two parts each lasting about one hour. In the first part, each subject completed a baseline task that measured absolute sensitivity at each of the four sounds against a background noise using an adaptive two-alternative, forced-choice (2AFC) task to ensure that thresholds were independent of criterion and all stimuli were equally detectable. Descending tracks were obtained from each subject using a 2-down, 1-up rule. Signal level was either increased by 1 dB after a single incorrect response or decreased by 1 dB after two consecutive correct responses. Thresholds corresponding to a probability of $0.5^{1/2}$ or 70.7% correct in 2AFC were obtained by averaging across 12 reversal points.

In the second part, separate probe signal tests were run for frequency and object number. For blocks of trials measuring selective attention to frequency, the frequencies of cue-target pairs were randomly selected from among four possible combinations: low-low, high-high, low-high, or high-low. For such blocks, object number was held constant within blocks at either "few" (4-6 objects) or "many" (23-32 objects). Similarly, cue-target pairs for blocks of trials for object number were randomly selected from among the following four sets of values: few-few, many-many, few-many, and many-few. Frequency was held constant within blocks at either a low (1.6, 1.7 kHz) or high (4.2 kHz) value. Measurements obtained from the first part of the experiment were used to calibrate target stimulus levels individually for each subject. Each trial began with a 625-ms cue at a level 15 dB above the subject's threshold for that sound (see Figure 1). That

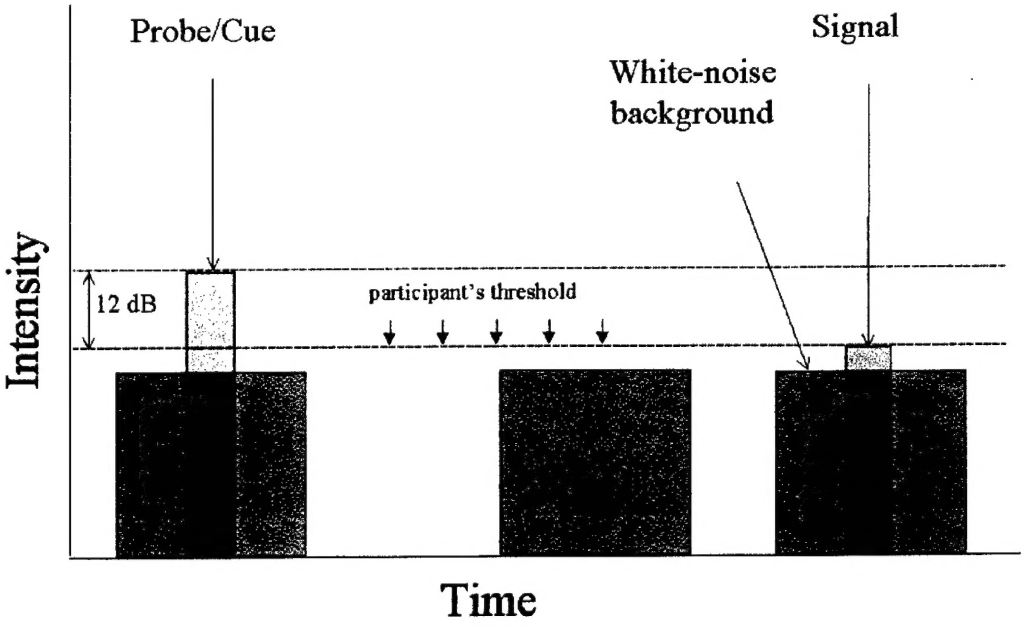



**Figure 1.** Time structure of each trial of the probe-signal paradim.

cue was followed after 1000-ms by two observation intervals separated by 500 ms, one of which contained a 625-ms target signal (probe) at the level of the subject's threshold: The cue and target were derived from the active "shaking" of the sound object and omitted the exponential decay that followed cessation of the shaking (a 50-ms artificial decay was superimposed at the end of the signal), to prevent subjects from attending simply to differences in decay rates. The subject then responded by indicating which interval contained the signal. Over the session, there was a 75% likelihood that the cue and target shared the same physical property (25% likelihood that the properties differed), and therefore cue validity was probabilistic rather than certain. Greater detectability with valid cues would imply the presence of attentional mechanisms associated with these object properties.

Stimuli were presented with the SigGen/PsychoSig psychoacoustic testing software packages operating in conjunction with Tucker-Davis Technology hardware and running on a Pentium microcomputer. Subjects listened to stimuli over Sennheiser HD265 headphones.

To compare detection performance across trials containing valid and invalid cues, for each subject we computed the proportions of correct responses for the two types of cues at each resonant frequency and object number. For the analyses of variance, we applied an arcsine transformation to the data to equalize variances. Figure 2a shows the untransformed data in Experiment 1 expressed as the percentage of correct responses for attended and unattended physical properties, shown separately for resonant frequency and object number [Bars for attended stimuli each comprise 720 judgments, whereas those for unattended stimuli comprise 240 judgments.] Detection performance was higher for attended (66.5%) vs. unattended (47.9%) targets generated by bamboo chimes of different resonant frequency [$F(1,9)=22.5$, $p=.001$], indicating that subjects were able to use object number in the detection task. There was no significant difference in detection performance for low (54.5%) vs. high (60.0%) resonant frequency cues pooled across attended and unattended targets [$F(1,9)=2.47$, $p=.15$].

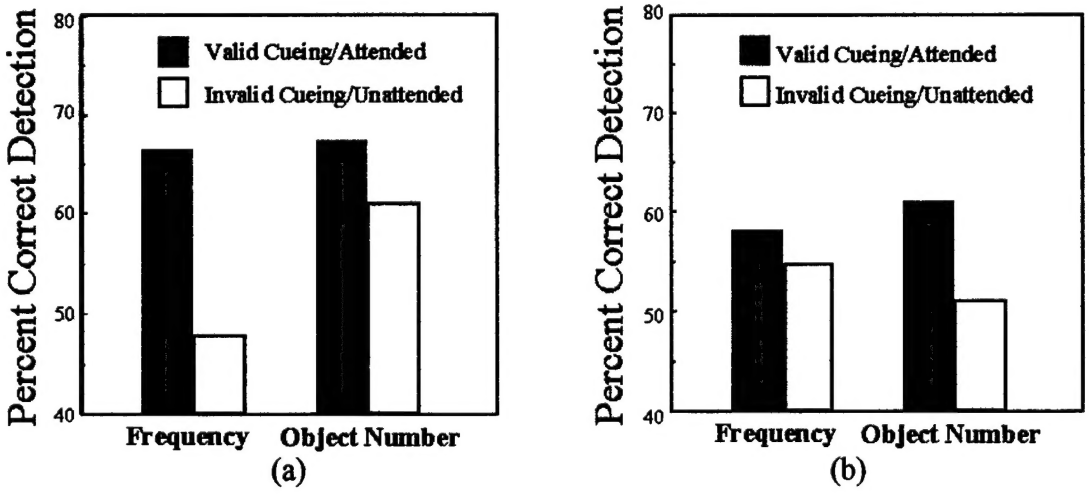



**Figure 2.** Listeners' detection performance. (a) When cue and target arose from the same object, listeners were able to use acoustic information on validly cued (attended) trials to attend to both resonant frequency and object number. (b) When cue and target came from different objects, selective attention was observed for object number, but not resonant frequency.

Similarly, when object number varied, detection performance was significantly higher for attended (67.5%) vs. unattended (61.2%) targets [$F(1,9)=6.67$, $p=.03$]. Curiously, detection was better for cues comprising many objects (67.2%) vs. cues with few (61.5%), although the difference misses statistical significance [$F(1,9)=3.8$, $p=.08$]. The reason for this latter performance difference is unclear, although we suspect that because the exponential decay function for sounds generated by many objects was less steep, a relatively longer segment of the signal may have been available at or near listeners' threshold for such sounds.

Results from both the resonant frequency and object number conditions indicate that listeners were able to attend selectively to these properties. The ability to attend to resonant frequency is perhaps not surprising given the considerable evidence for selective attention to pure tone frequency. However, selective attention to the number of colliding objects points to attentional mechanisms more complex than those found for pure tone frequencies, perhaps arising at central levels. In this light, we attempted to increase the complexity of the detection task in Experiment 2 by determining if selective attention to resonant frequency and object number would persist even if cue and target sounds arose from *different* percussive instrument models, which nonetheless shared these two underlying physical parameters.

### 4.1.3 Experiment 2

The ten subjects from Experiment 1 were also invited to participate in Experiment 2. All agreed to return and were compensated for their participation.

Stimuli were the four bamboo chime sounds used in Experiment 1 and four guiro sounds (see Figures 3e, f, g, and h) generated by crossing two levels of resonant frequency - "low" (2.1 kHz) and "high" (4.2-4.5 kHz) - with two levels of object number - "few" (4-6 objects) and "many" (28-32 objects). Although the spectral centroids of the low frequency guiro sounds were higher than those of the low frequency bamboo chimes - a function of differences in the physical models appropriate to simulating these two percussive instruments - we were confident that they would nonetheless serve as effective cues to resonant frequency.

Thresholds for the four guiro sounds were obtained for each subject with the tracking method of Experiment 1. The probe signal paradigm was also retained, except that cue and target on each trial came from different instruments: cue (guiro)-> target (bamboo chimes), or cue (bamboo chimes)->target (guiro). Software/hardware were retained from Experiment 1.

The untransformed percentage of correct responses for attended and unattended physical properties is shown separately for resonant frequency and object number in Figure 3b. With different percussive instruments for cue and target, listeners showed no significant difference in detection performance for resonant frequency for attended (58.2%) vs. unattended (55.0%) targets [$F(1,9]=.01$, $p=.92$], contrary to our initial expectations. Perhaps information about resonant frequency cannot be used to direct attentional mechanisms to sounds with different sources and/or highly distinct timbres. Detection of low (56.4%) vs. high (56.8%) resonant frequency cues pooled across attended and unattended targets was not significantly different [$F(1,9) = .233$, $p=.64$]. Listeners were

better able to detect attended (61.1%) vs. unattended (51.4%) targets when cued to the number of colliding objects [F(1,9)=5.19, p=.05], suggesting that the perceptual salience of this parameter transcended the particular physical model (i.e., bamboo chimes or guiro) that it controlled. Detection was no better for cues comprising many objects (54.7%) vs. those comprising few objects (59.7%), [F(1,9)=.41, p=.54].

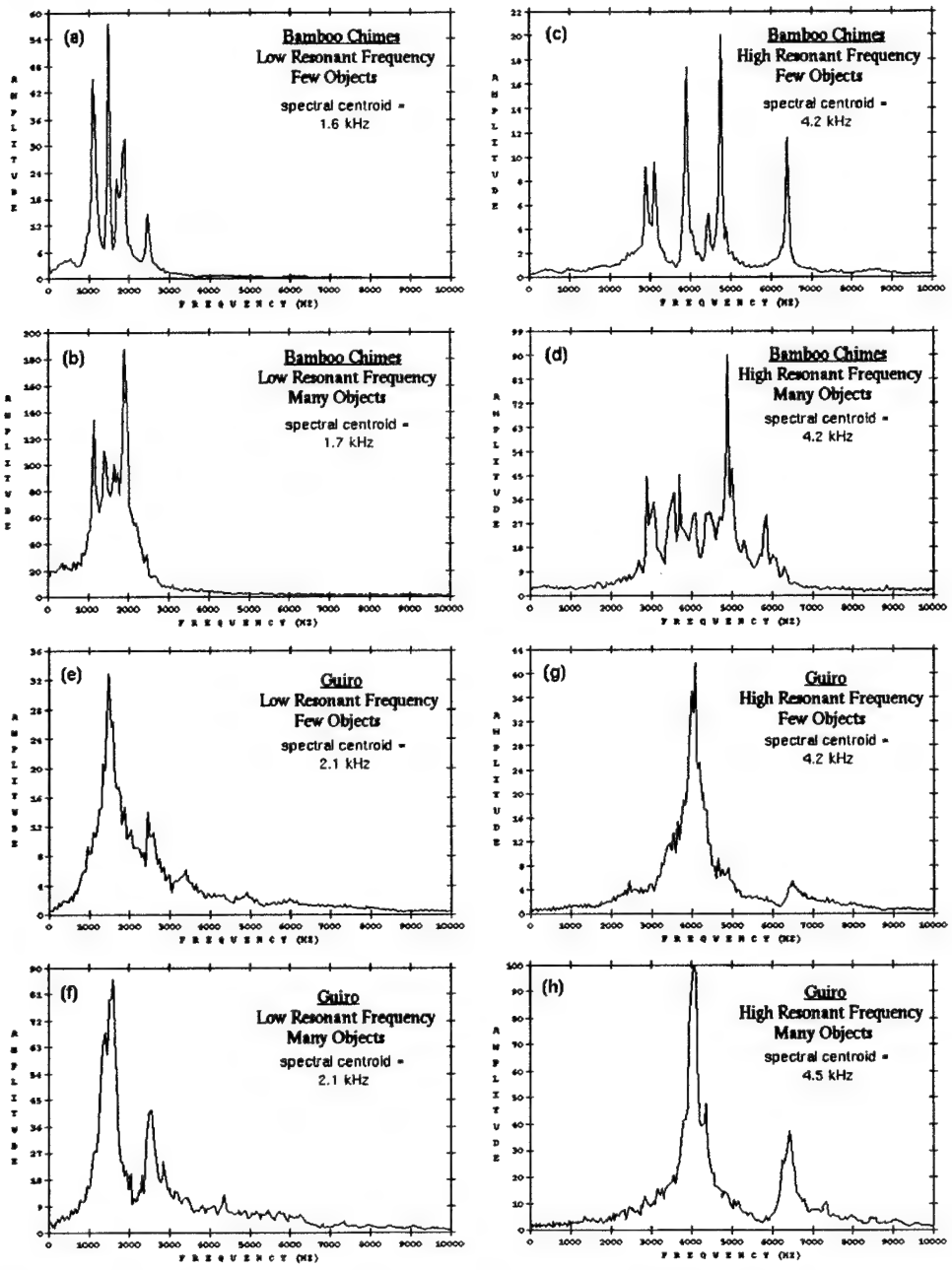



**Figure 3.** Spectra for the four bamboo chime stimuli in Experiments 1 and 2, and the four guiro stimuli in Experiment 2. Each set of four stimuli was generated by crossing two levels of resonant frequency - "low" (1.6-1.7 kHz centroid for bamboo chimes, 2.1 kHz for guiro) and "high" (4.2 kHz for bamboo chimes; 4.2-4.5 kHz for guiro) - with two levels of object or particle number - "few" (4-6 objects) and "many" (28-32 objects).

The results point to a fairly remarkable ability of the auditory system to monitor individual acoustical properties of sound sources. Perhaps most impressive was that subjects were still able to attend selectively to the number of objects in Experiment 2, even when the cue and target signals were generated by different objects. Conversely, the absence of selective attention to resonant frequency in this context may point to limitations to frequency-based attention across widely disparate timbres. It remains unclear precisely what features of the proximal waveform may be the most salient "markers" of these distal acoustic properties. However, given the quasi-random nature of the collisions simulated by the PhISM models used here and the ability of listeners to use information about properties in one instrument to attend to those in another, it is clear that the proximal spectral and temporal cues must be highly complex. In future work, we intend to isolate classes of such cues in a number of different physical models to derive a predictive, ecological model of timbre that links the acoustic properties of sounds and their sources to their perceptual correlates.

## 4.2 Acquiring Knowledge about Sound Source Properties

**SUMMARY:** The study's goal was to assess how listeners acquire knowledge, or learn, about sound sources. We had three specific aims for the current study: (1) to measure how listeners learn about the physical parameters of sound sources over time, (2) to measure how the amount of structure in the learning environment may modulate learning about sound sources (3) to introduce both perceptual and cognitive measures of learning since there may be many types of learning going on at multiple levels of processing.

### 4.2.1 Method

Seventy-five subjects were recruited from Washington State University Vancouver for this study and compensated for their participation. There were 15 subjects recruited per condition. Most subjects had little or no musical training so could be collectively called "naïve" to issues relating to physics of musical/percussive instruments.

Each subject completed three alternating phases of learning and testing. Listeners first engaged in a 15-minute learning session in which they explored an interface for the shakers algorithm; they then completed a 15-20 minute testing session, and then they completed two more learning-testing pairs for a total of three. The amount of structure in the shakers interface was varied across four subject groups; a control group completed only the testing phases with 15-minute breaks where the others were taking part in the learning sessions. The testing phases comprised three components: a discrimination task, a memory task, and a similarity-rating task, and I will describe each of these in a moment. And the amount of learning by listeners in the different groups was assessed by performance changes across the three testing phases.

Figure 4 illustrates the interfaces for the four learning conditions. For all four interfaces, eight instruments were tested: maraca, cabasa, guiro, bamboo wind chimes, tambourine, sleight bells, ratchet, and a coke can, designed by Dr. Perry Cook to sound like a can being dragged along a rough surface. Listeners were permitted to explore as they wished, but were provided some general guidelines concerning roughly how frequently they should move the sliders and how they should explore all of the sliders. In the "highly structured" learning condition (Figure 4a), the instruments were labeled and illustrated and subject were provided with brief written descriptions of the instruments taken from music dictionaries. Sliders were grouped by instrument, and they were ordered and labeled. Figure 4b shows the "moderately structured" condition, in which no pictures or labels were given for the instruments, but in which the sliders were grouped by instrument and they were ordered and labeled. In the "weakly structured" condition Figure 4c), sliders were grouped according to instrument, but slider order was randomized within each instrument. And finally, in the "unstructured" condition (Figure 4d), the orders of the 32 sliders were randomly positioned and not grouped according to instrument.

For the discrimination task, subjects judged whether two sounds presented in sequence were the same or different. The pair of sounds always came from the same instrument - which could be the maraca, tambourine, sleigh bells and coke can - but could vary on any given trial along one of two physical parameters - damping and object number. In pilot testing, resonant frequency had turned out to be too easy a task, with listeners performing

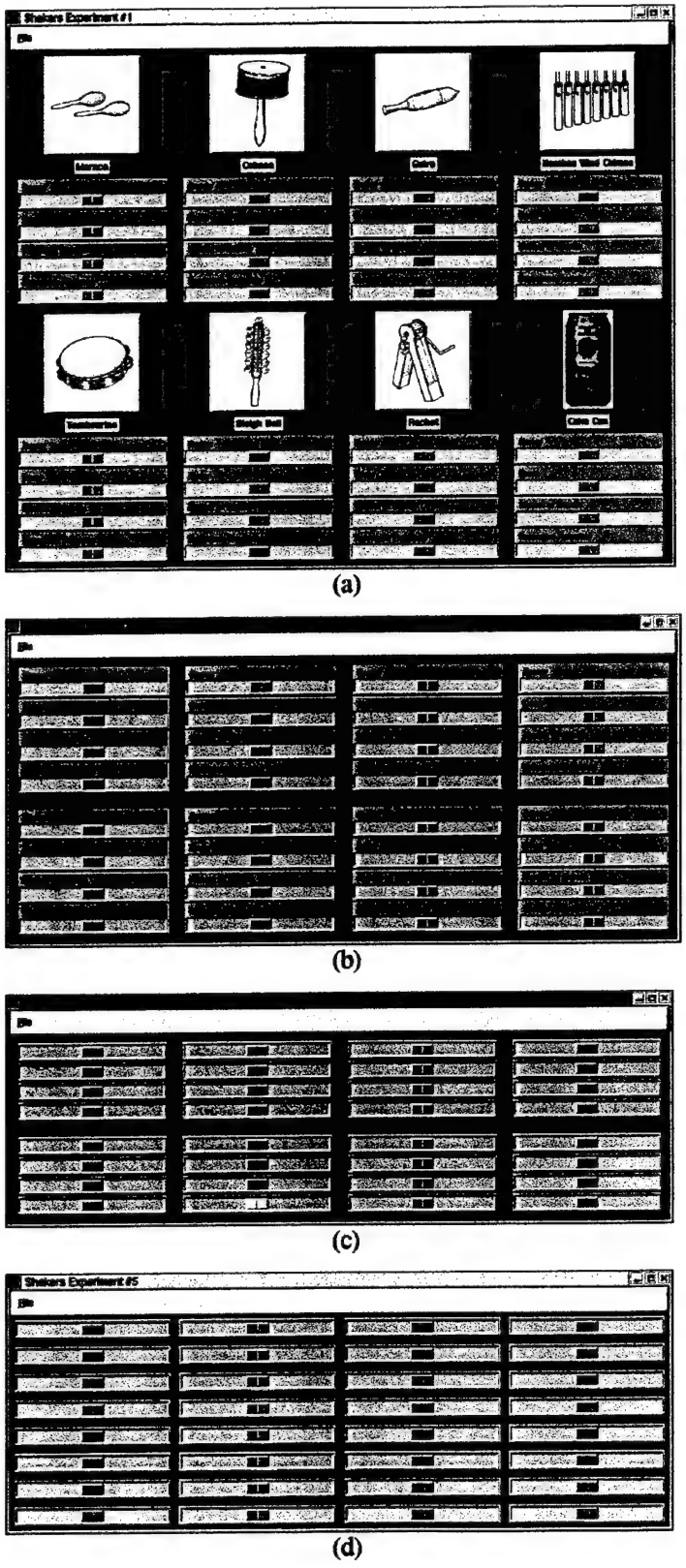


**Figure 4.** The four graphical interfaces for the shaker instruments used in the learning phase: (a) highly structured; (b) moderately structured; (c) weakly structured; (d) unstructured.

at ceiling. The four instruments selected had what we thought were roughly equal variations in step size along the damping and object number sliders. For some instruments - the ratchet for example - the number of objects slider does not induce much change in the sound.

For the memory task, listeners determined whether or not a target presented after a sequence of sounds is present in the sequence; the length of the sequence varies between three and seven sounds. For this task, all the eight shaker instruments were tested, and the physical parameters were held constant within each instrument across trials; the memory task, therefore, measures listeners' ability to remember source identity. Trial-by-trial feedback was provided for both the discrimination and memory tasks.

For the similarity-rating task, subjects estimated the similarity for all pairwise comparisons of 12 sounds. The four instruments from the discrimination task were tested, with variations introduced in each instrument for resonant frequency, damping and object number. For each of these three physical parameters, we used the left and right endpoints of the step series we had generated for the discrimination task, so parameters are clearly discriminable. Participants used a 9-point rating scale to estimate similarity. Similarity matrices were submitted to a multidimensional scaling program (e.g., CLASCAL: Winsberg & DeSoete, 1995) to isolate differences in the structure of perceptual spaces between blind and sighted participants.

A control group completed only the testing phases with 15-minute breaks where the others were taking part in the learning sessions. Amount of learning across the different participant groups was assessed by performance changes across the three testing phases. The testing and learning session each lasted 15 minutes; the duration of the testing session varied depending on the speed of the individual listener, but was typically between 15 and 18 minutes. The total duration of the experiment was approximately 2 hours; subjects complained a lot less. Task order was randomized within testing sessions. Listeners were tested in a double walled IAC booth and the presentation of sounds was accomplished using a TDT System 2 and Sennheiser HD 265 headphones.

### 4.2.2 Results

For the results from the discrimination task, only the number of objects parameter, and not the damping parameter, showed differential effects, so the discussion will focus on results pertaining to the number of objects parameter. Figure 5 plots discrimination performance across testing sessions 1 and 3 as a function of the degree of difference in number of objects across the pair of sounds on each trial. One sound of the pair always had a fixed, intermediate number of objects while the second one had either fewer or more objects composing it. Listeners tended to respond same 75% of the time even though the probability of the sounds being the same was 50% across the session; listeners' criteria in this respect were not significantly different across the four conditions as well as the control condition, so they are plotted them on the same graph. Performance improved as one of the sounds of the pair had fewer or greater number of objects that the other, but there was no significant difference in performance across the four learning conditions after learning session 1. Also, the performance in the four conditions was no different from performance in the control.

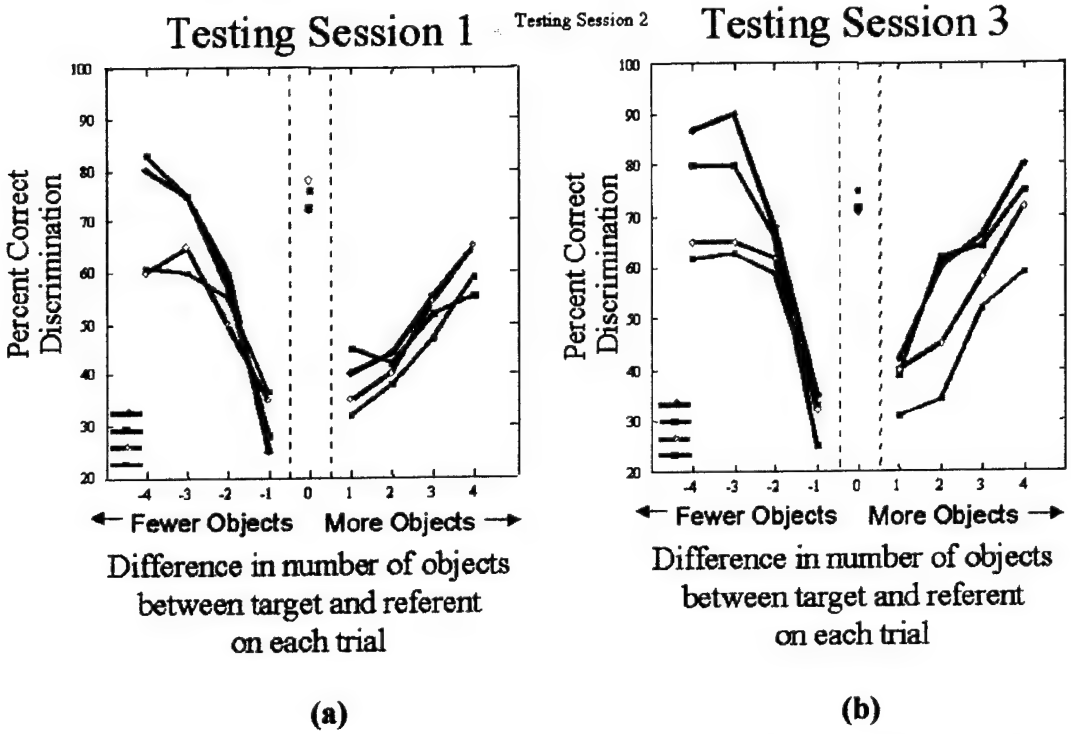



**Figure 5.** Discrimination scores are shown as a function of the difference between the number of objects composing each of two shaker instruments presented on each trial. Figures 3a and 3b display discrimination performance in testing sessions 1 and 3, respectively. Separate functions are shown for the four conditions that varied the amount of structured information available through the computer interface to the shaker instruments. Improvements in discrimination are proportional to the amount of structure presented in the learning interface.

By the third testing session, however, there are significant differences in performance across conditions, even weakly structured and unstructured ones. Performance for all conditions except control in testing session 3 is significantly better than in session 1. The magnitudes of the improvements are linked to the amount of structure in the learning condition the listeners were in. The magnitude of improvement is positively correlated with the degree of structure in the learning condition the listener is in. There was no significant improvement for the control group, however. It is interesting to note that the learning effect is more pronounced for discriminations involving greater numbers of objects. This may simply be due to the fact that initial JND's were in fact not quite equal in size, but it could also mean that listeners may initially know more about small numbers of objects from simple experience but come to learn about discriminating larger numbers of objects here in the learning conditions.

Figure 6 plots results from the memory task, comparing target identification performance across sessions 1 and 3 for the four learning conditions, as well as the control. As expected, performance declines with increasing numbers of sounds in the sequence in which the target has to be detected. In testing session 1, there are no significant differences between identification rates across the conditions. In session 3, however, the two most structured conditions are the only ones in which performance improves significantly. In other words, performance decline for larger numbers of items in the sound sequences, but improves significantly across testing sessions 1 and 3 only for the

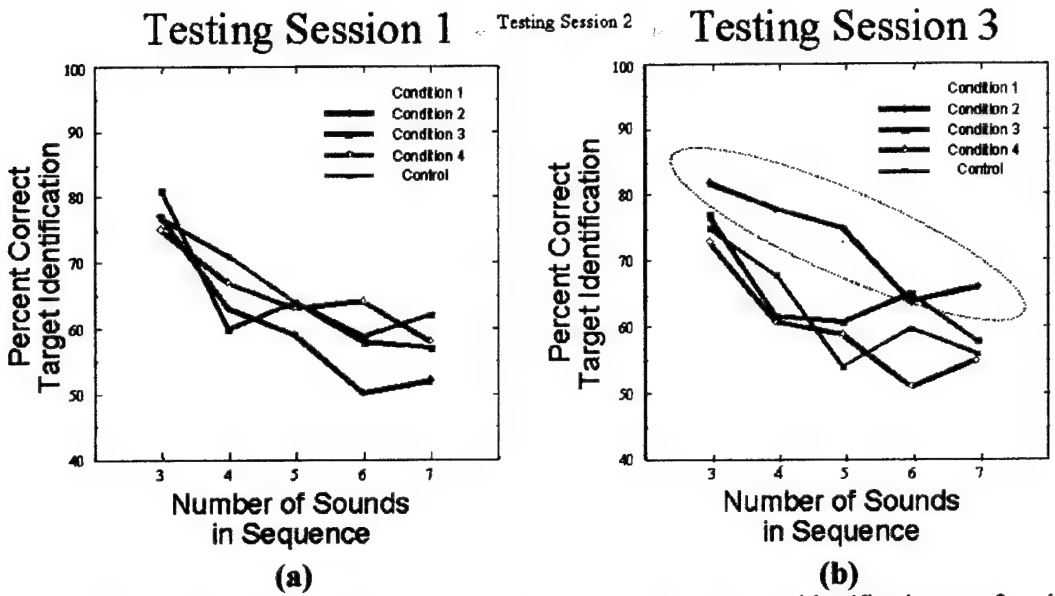



**Figure 6**. Results from memory task, expressed as percent correct target identification as a function of the number of sounds presented in the test sequence. Figures 4a and 4b display discrimination performance in testing sessions 1 and 3, respectively. Separate functions are shown for the four conditions that varied the amount of structured information available through the computer interface to the shaker instruments. Significant improvements in identification occur only for the two most highly structured learning conditions (Conditions 1 and 2).

highly and moderately structured conditions, suggesting that for the memory task, the amount of structure in the learning session may be more critical than for the discrimination task.

The results for the similarity rating task are somewhat equivocal, although analyses are still ongoing. Although the optimal solutions based on Kruskal stress measures for the more highly structured conditions have more dimensions in general, it's not really clear that there is are strong trends, either across conditions or across learning sessions.

Given the exploratory nature of the overall study, perhaps the most important finding it conveys that that heuristics-based physical models can be very efficient and practical tools for study auditory source perception. Although exhaustive models can offer a greater degree of precision, they are difficult to develop; the flexibility in stimulus control inherent in the shaker model can allow us to understand some global aspects of source perception in an elegant way.

## 4.3 An Interactive Computer Environment for Obtaining Perceptual Spaces for Large Numbers of Complex Sounds

**SUMMARY:** The Sonic Mapper is an interactive Linux-based graphical program that affords increased methodological flexibility and sophistication to researchers who collect proximity data for auditory research. The Sonic Mapper consists of a mapping environment in which participants can position and group icons in the two-dimensional plane of the screen. Options for collecting data concerning hierarchical groupings, category prototypicality, and verbal labeling provide additional opportunities to test hypotheses in a convergent manner. The Sonic Mapper also offers an environment for traditional pairwise comparisons, as well as one for performing free sorting tasks. A pilot study that attempts to use many of the Sonic Mapper's key features is described briefly below.

### 4.3.1 Background

Conventional multidimensional scaling typically requires participants to provide similarity ratings for all possible stimulus pairings. For a complete data matrix, this yields a total number of comparisons as computed using Equation 1,

$$C = \frac{N(N-1)}{2} \qquad (1)$$

where $C$ is the total number of comparisons to be made, and $N$ is the number of stimuli.

With even 50 stimuli, for example, participants are required to perform 1225 judgments. Aside from obvious considerations of fatigue, it is questionable whether participants can maintain a consistent set of criteria across so many comparisons. Primarily designed to address the issues of fatigue and criteria inconsistency associated with the collection of similarity ratings for large stimulus sets, the Sonic Mapper is an interactive Linux-based program through which listeners arrange sound stimuli in a manner that reflects their relative similarities. In addition to addressing the issues for which it was originally designed, the Sonic Mapper also provides some distinct advantages over conventional pairwise comparisons.

Constraining participants to make pairwise comparisons, for instance, may discourage them from generating and applying richer sets of perceptual and cognitive criteria in their ratings. For instance, if sounds A and B have attributes v, w, and x in common, and sounds A and C have attributes x, y, and z in common, it is less likely that a participant performing pairwise comparisons will discover that B and C have attribute x in common than if he or she is able compare A, B, and C simultaneously. The Sonic Mapper provides a visual display of how all sounds relate to one another simultaneously. In addition, allowing participants to group and colorize sounds with common attributes facilitates the identification of these attributes, especially over extensive testing. Static visual display of the stimuli (via icons) alone, although it may enhance participants' global appreciation of the stimulus set, does not allow participants to select which, and in what order, items are to be compared. Such control of comparisons is afforded by the interactive nature of the Sonic Mapper.

In a related context, conventional randomized presentation of pairwise comparisons makes it impossible for participants to exert control over the order of comparisons - to adjust their similarity rating, for example, for a previously presented pair in light of a new criterion that he or she may have generated on the basis of a later comparison – and thus may discourage the development of alternative decisional strategies. The interactivity of the Sonic Mapper leaves decisions regarding how stimuli are compared primarily to the participants. They are free to employ any strategy and change strategies at any time throughout the experimental process. We believe this interactivity also allows participants to develop a more global appreciation of the stimuli than they could achieve through pairwise comparisons alone. However, such interactivity reduces the amount of structure to the experimentation. For this reason, the Sonic Mapper can also enforce specific comparisons. The program can either read these comparisons from an external file or automatically generate a set of comparisons based on configurable parameters such as minimum comparisons per item and percentage of all possible comparisons. This feature ensures that listeners make certain comparisons in a pairwise manner, and yet, it still allows participants the freedom to choose how, and in what order, to approach these comparisons.

The predominant use of continuous rating scales for pairwise comparisons may encourage participants to think unidimensionally about the similarities between stimulus pairs. For example, when dissimilarities between two sounds are represented along a unidimensional slider, listeners are less likely to consider both amplitude and pitch when rating sounds. Instead, listeners can, and often do, make their ratings based on the more salient attributes of the sounds (i.e., the attribute with the greatest similarity), or attempt some "cognitive averaging" that is unlikely to remain consistent over multiple pairs and obscures the attributes from which the averages were derived. The Sonic Mapper allows for similarity ratings along two continuous dimensions represented by the Euclidean distance between individual icons. Although two-dimensional ratings may appear advantageous over the unidimensional scales employed in conventional pairwise comparisons, the employment of 2-D representation of distance limits the listeners' ability to rate beyond these two dimensions – to think outside the plane of the screen, so to speak. This limitation, however, is, one that the Sonic Mapper shares with pairwise comparisons; there is no evidence that participants will rate stimuli beyond the single dimension provided to them when using a unidimensional rating scale.

The majority of published MDS solutions seem to be two-dimensional, and it is indeed rare to see solutions in the literature with four or more interpretable dimensions. It is unclear whether this may be a by-product of MDS algorithms themselves (i.e., that they tend to generate optimal solutions that are low-dimensional) or whether participants cannot access more than two or three perceptual/cognitive dimensions simultaneously, even though more might potentially be available. If low-dimensional solutions predominate, we wonder whether it might be better to make this assumption explicit in an algorithm, and to treat the presence of higher dimensions as an exception, rather than the rule. Finally, because of the two-dimensional constraint imposed by the computer screen, continuous similarity ratings made using the Sonic Mapper must be restricted to two dimensions.

### 4.3.2 Overview of Sonic Mapper Features

Sonic Mapper allows participants to provide similarity estimates for large stimulus sets in an efficient manner. Participants move icons corresponding to individual sounds from a docket into the two-dimensional space of the computer screen in order to reflect relative similarities within the sound set. Afterwards, the Euclidean distances for the full matrix of sounds are computed. The experimenter can set several limits on the number and nature of comparisons per stimulus that participants must make during a session. Mnemonic devices are included to allow participants to keep track graphically of the nature and extent of past comparisons. There are several features included in the Sonic Mapper that augment the experimenter's ability to detect situations in which a k-D ($k \geq 3$) solution is more appropriate.

Sonic Mapper provides a high degree of interactivity, including opportunities for multiple comparison strategies, availability of mnemonic cues to remind participants of past comparisons, and the presence of feedback concerning classification schemes. SonicMapper allows a participant to form multiple hierarchical levels of groupings and subgroupings of stimuli using colorization and boundary definitions, with self-generated labels for each group (see Figure 7).

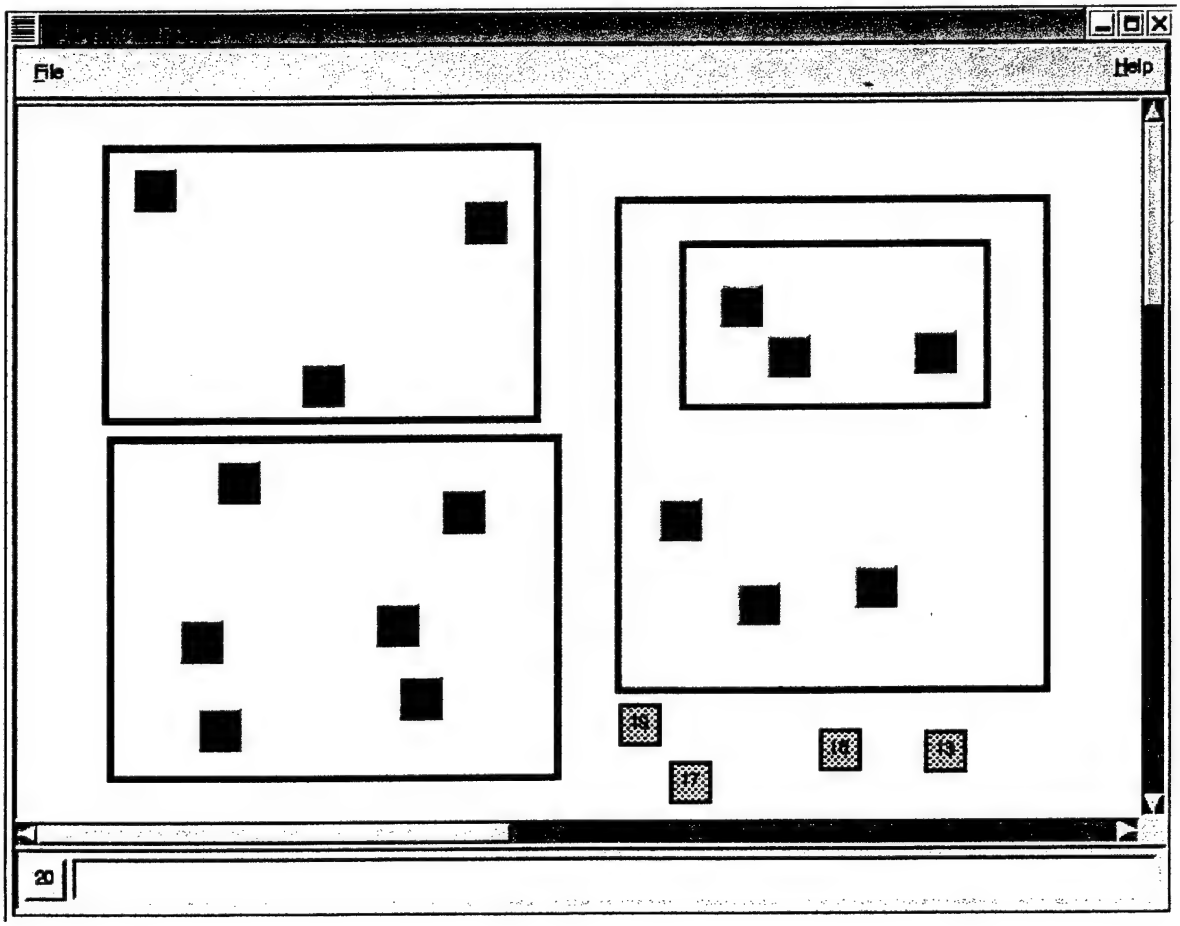



**Figure 7.** Main mapping window permits the participant to generate a two-dimensional similarity space with the option of superimposed hierarchical grouping boundaries.

The experimenter can define a number of parameters separately for each participant before running a sound mapping session. In addition to allowing the specification of stimulus sets and basic display features, the program provides the option of requiring pairwise comparisons – derived randomly or according to a predefined matrix – on a desired subset of the stimulus set (Figure 8). The required comparisons feature permits one to test the validity of the mapper's two-dimensional layout by deriving a second multidimensional solution from the same data set based on the pairwise comparisons. An optional backup feature saves the current state of a participant's mapper configuration at specified intervals; if states are saved to separate files, a "time-lapse" view of a participant's arrangement of icons, consisting of an arbitrary number of sequential snapshots, can be obtained across the course of a complete experimental session.

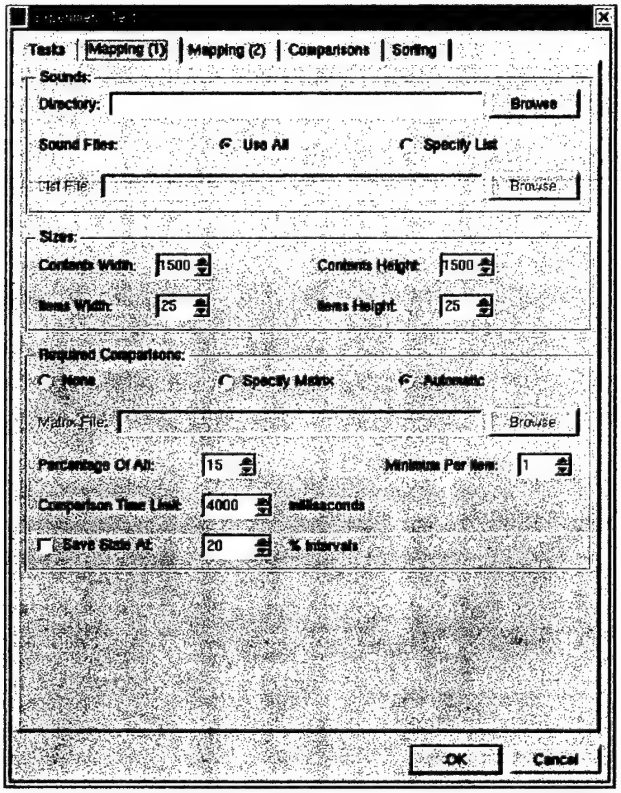



**Figure 8.** The first experimental specification window for the mapping task. Sound stimuli, visual display parameters, and required pairwise comparisons can be specified here.

Three supplementary data collection options allow the experimenter to test additional hypotheses about dimensionality and categorization (see Figure 9). After a participant completes his or her stimulus map, the experimenter has the option of requesting confidence ratings for the final location of each stimulus. Stimuli with low confidence ratings can then be tested further to determine whether they signal the need for additional dimensions or whether they represent "outliers" that have unique properties not easily represented by common dimensions. An option for permitting hierarchical groupings can also be enabled by the experimenter. Upper and lower boundaries can be placed on the number of groups or categories a participant is allowed to generate. Each group can be colored and labeled by the participant (labeling can either be forced or left optional by the experimenter). A separate menu item also allows the participant to view the entire

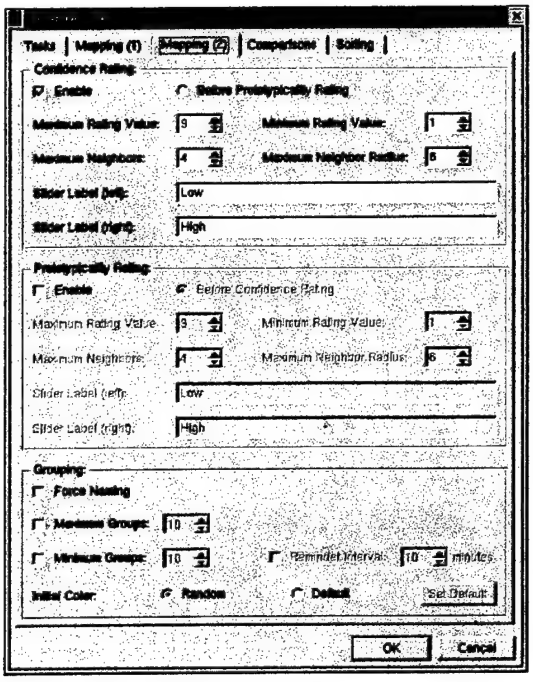



**Figure 9.** The second specification window for the mapping task. Control is provided over several options for data collection, including hierarchical grouping, confidence and prototypicality ratings.

grouping scheme for the mapper screen in a graphical tree format. For those interested in obtaining more specific information about group membership, Sonic Mapper provides the option of collecting prototypicality ratings for each stimulus relative to its membership.

To provide additional opportunities for collecting alternative comparison data within a single experimental session, Sonic Mapper has utilities for obtaining proximity data in the form of pairwise similarity ratings as well as ordinal sorting data. Figure 10 illustrates

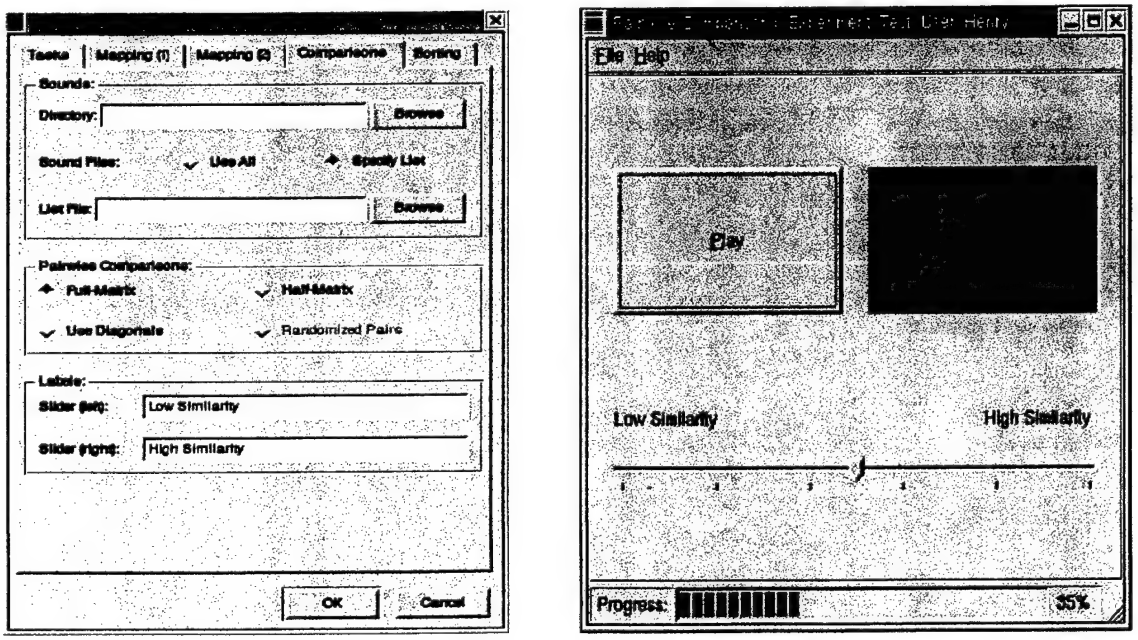



**Figure 10.** Sonic Mapper windows for pairwise comparison task.

the interface used to collect pairwise similarity data. Participants adjust a continuously variable slider whose poles can be labeled by the experimenter in a separate specification window. The experimenter can also specify whether a full or half matrix is to be used to generate stimulus pairs (for half- matrix usage, an option to include or exclude the matrix diagonal is provided). Sound stimuli may be obtained by directing Sonic Mapper to a particular directory or to a text-file containing a list of the sounds' names.

For the sorting task, sound stimuli are presented sequentially when the participant clicks on the speaker icon at the bottom of the main sorting window; stimuli are then place in one of $n$ folders or bins specified by the experimenter. Through a separate specification window (see Figure 11a), the experimenter can place upper and lower bounds on the number of sorting groups or "bins" that the participant is permitted to create. Sorting groups comprise folders whose contents can be accessed at any time during an experimental session (see Figure 11b), so that the participant may reallocate stimuli to other groups at any time. Verbal labels may be provided for the groups by either the participants or, *a priori*, by the experimenter.

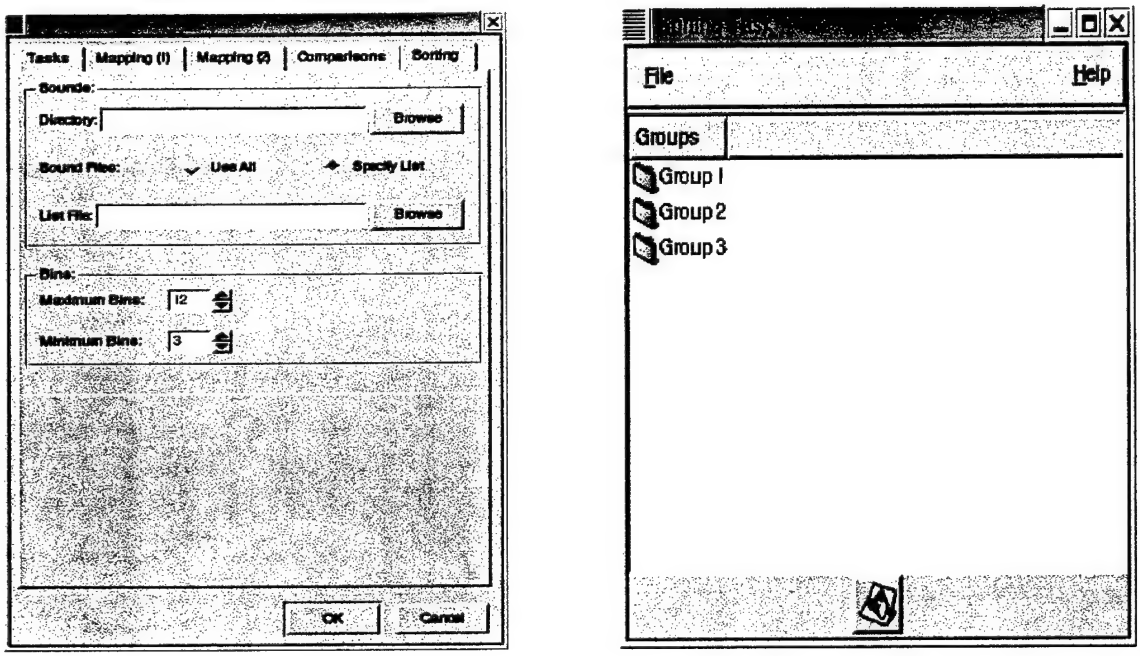



**Figure 11.** Sonic Mapper windows for sorting task.

### 4.3.3 Pilot Testing of the Sonic Mapper

In our first use of Sonic Mapper, we obtained similarity judgments for 150 complex sounds from 22 participants. Stimuli were sound effects recorded from a variety of effects libraries (e.g., BBC). The sound effects all involve human-object interactions – a person walking on gravel, a person typing on a typewriter, or someone clinking a cup and saucer together. Since most of these stimuli evoke strong mental images of the sources or objects generating them, we compared participants who were instructed to focus on the timbre of each stimulus with participants who focused on the mental image generated by each stimulus. Participants were asked to generate between 5 and 15 hierarchical stimulus groupings using the interface, and participants in the mental imagery condition were also asked to provide verbal descriptors for each stimulus. In order to test the

validity of a two-dimensional assumption for this space - since Sonic Mapper allows arrangement of stimuli in two dimensions only - we subsequently obtained pairwise similarity ratings from each participant for all within-category comparisons, as well as a randomly selected subset of across- category comparisons, and submitted the resulting dissimilarity matrices to MDS analysis.

Figure 12 shows the MDS solution for pooled similarity matrices for the "mental imagery" condition. In most cases, items clearly tend to group according to shared physical properties (e.g., "door" sounds at top of figure, "walking" sounds at middle right), in contrast to the timbre condition (figure not shown), where source-based clustering is not apparent. One general observation regarding how participants create their spatial arrangement: We were surprised by the enthusiasm that participants showed for the mapper task, working for several hours at a stretch. Most seemed to have a vested

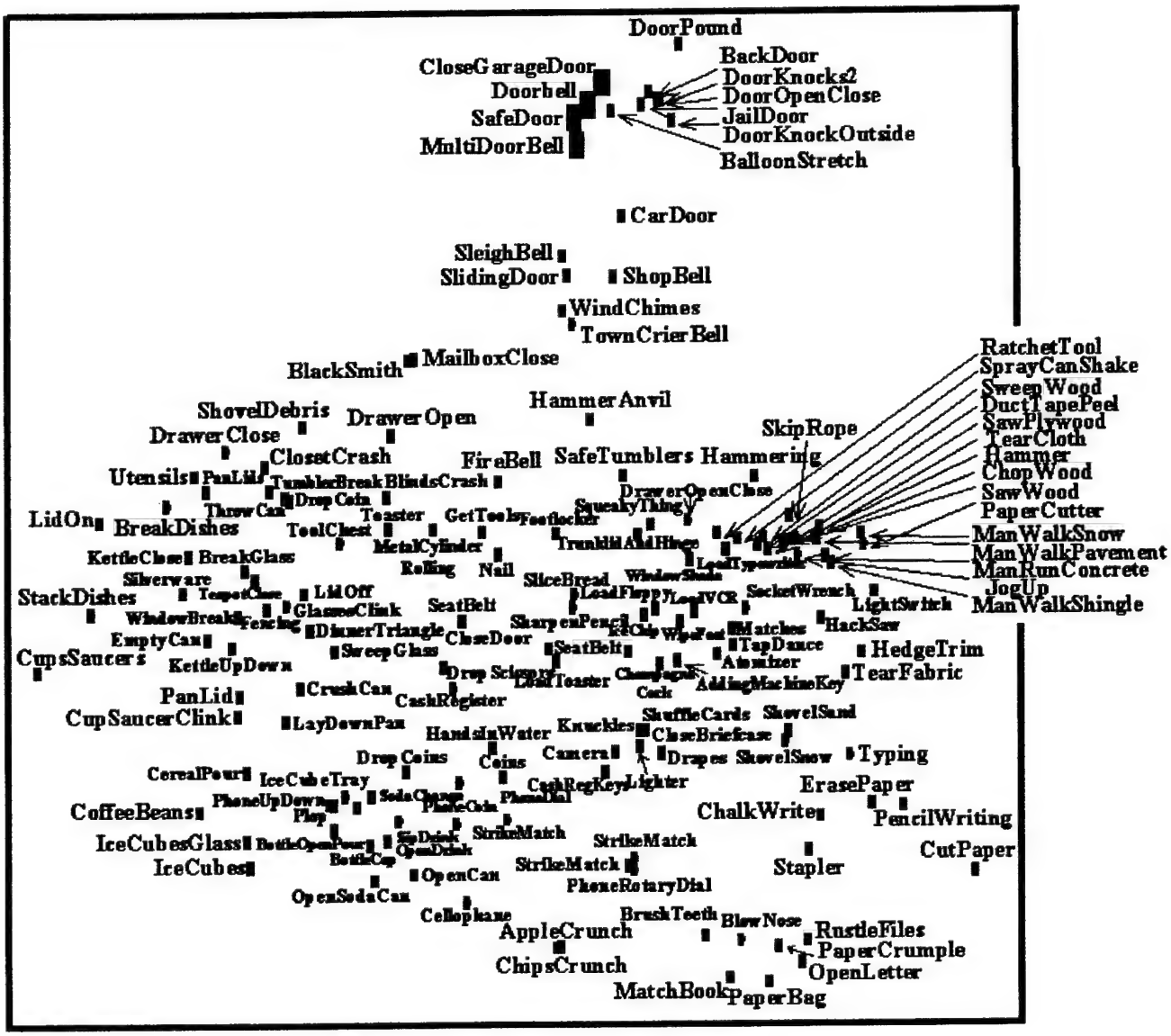



**Figure 12.** Multidimensional scaling solution (CLASCAL) derived from pooled similarity matrices obtained using Sonic Mapper.

interest in thinking through carefully the complex problem of organizing large numbers of stimuli into a coherent perceptual representation, and many seemed genuinely proud of what they accomplished by the end. We conducted tape-recorded interviews with participants at the conclusion of the experiment, and were impressed with their ability to describe in clear terms the global aspects of their perceptual space. In short, the interactive character of the Sonic Mapper seemed to turn the task into an engaging problem-solving one.

In sum, the Sonic Mapper provides a viable alternative to conventional pairwise comparisons for collecting similarity data. In addition to being appropriate for large stimulus sets, the Sonic Mapper provides simultaneous presentation and manipulation (interactivity), as well as a two-dimensional rating scale. However, the Sonic Mapper is not without its limitations. Use of the Sonic Mapper in situations that require more than two dimensions is questionable, although still possible using an ordinal hierarchical sorting task, albeit with a loss in precision. Classification and grouping of stimuli along with confidence ratings are also incorporated in the Sonic Mapper's methodological repertoire to help the experimenter detect situations where a *k*-D solution would be most appropriate. Future versions of the Sonic Mapper will allow for visual stimuli sets and cross-platform use.

## 4.4 Extended Perceptual Spaces for Harmonic and Percussive Timbres

**SUMMARY:** MDS studies of auditory timbre have usually yielded two- or three-dimensional solutions, and at least two perceptual dimensions appear consistently: (1) The first perceptual dimension correlates strongly with the "spectral center of gravity" or spectral centroid of a sound; (2) The second dimension frequently correlates with the rise time or attack of the sound. As past studies conducted in the principal investigator's laboratory have shown, these dimensions are perceptually primary even for highly diverse stimulus sets. One possible explanation for such results is that spectral centroid and attack time are physical correlates of such strong perceptual dimensions that even if higher dimensions potentially exist, these first two dimensions will tend to mask them. A series of listening tests attempted to better isolate higher-order perceptual dimensions of timbre by building on past findings showing that spectral centroid and rise time represent principal acoustic correlates of the primary timbral dimensions. Listeners rated the timbral similarity of three sets of pitched and percussive instrument sounds that were equated for centroid and rise time using signal-processing techniques. Multidimensional scaling analyses yielded two- and three-dimensional perceptual spaces for the three stimulus sets. Several aspects of the spectral fine structure of the timbres – including spectral irregularity, incoherence, and density - correlated with the dimensions of the three spaces, as did signal decay. The results indicate that additional spectral and temporal cues are available to listeners when the influence of spectral centroid and rise time is removed. However, the frequent presence of multiple acoustic correlates for individual dimensions suggests that it may be more ecologically appropriate to interpret such correlates in the context of general physical properties of the instruments themselves.

### 4.4.1 Method

Sixteen women and 9 men between the ages of 18-40 served as participants. All were undergraduates at Washington State University recruited through posters throughout the campus, and none had participated in any prior auditory experiments. Participants were compensated $20 for approximately two hours of experimentation.

Stimuli were modified versions of 35 sounds used by Lakatos (2000) and selected from the McGill University Master Samples (MUMS) compact discs (Opolko and Wapnick, 1987) of digitally recorded musical instruments, as well as an additional recorded sound (tam-tam), bringing the total stimulus set to 36. As, shown in Table 1, stimuli were originally selected to fall into two general sets of 18 sounds each: Sounds produced by traditional pitched orchestral instruments (e.g., flute, trumpet, piano) and sound produced by traditional and non-Western percussive instruments (e.g., bamboo chimes, tubular bells, cuica). We maintain the nomenclature from Lakatos (2000) by referring to these two stimulus sets as the "harmonic set" and "percussive set," respectively. A third set, hereafter referred to as the "combined set", was generated by selecting 10 sounds each from the harmonic and percussive sets. Additional details concerning the criteria used for the selection and preparation of this original stimulus set can be found in Lakatos (2000). All three stimulus sets were then modified through signal processing techniques described below in order to normalize their spectral centroids and rise times.

| Harmonic | Percussive | Combined |
|---|---|---|
| • Baroque Recorder | • Bamboo Chimes | • Baroque Recorder |
| • Bb Clarinet | • Bongo Drum | • Bb Clarinet |
| • Tenor Crumhorn | • Castanets | • Flute (flutter-tongued) |
| • English Horn | • Celesta | • French Horn |
| • Flute (no vibrato) | • Cuica | • Harp |
| • Flute (flutter-tongued) | • Cymbals (bowed) | • Harpsichord |
| • French Horn | • Cymbals (struck) | • Piano |
| • Harp | • Log Drum | • Tenor Saxophone (growls) |
| • Harpsichord | • Marimba | • C Trumpet (muted) |
| • Piano | • Snare Drum | • Violin (martelé) |
| • Pipe Organ | • Steel Drum | • Celesta |
| • Alto Saxophone | • Tambourine (pop) | • Cuica |
| • Tenor Saxophone (growls) | • Tam-Tam | • Cymbals (bowed) |
| • Bb Trumpet | • Temple Block | • Cymbals (struck) |
| • C Trumpet (muted) | • Tubular Bells | • Log Drum |
| • Violin (no vibrato) | • Tympani | • Snare Drum |
| • Violin (martelé) | • Vibraphone (bowed) | • Steel Drum |
| | • Vibraphone (struck) | • Tubular Bells |
| | | • Tympani |
| | | •Vibraphone (bowed) |

**Table 1.** Thirty-six musical instruments used as stimuli in three stimulus sets.

Each stimulus sound was first analyzed using a phase vocoder program [Beauchamp, 1992]. This program assumes that the input sound can be represented as

$$s(j) = \sum_{k=1}^{K} A_k(j)\cos(2\pi\int(kf_a + \Delta f(j)) + \theta_k(0)) \qquad (1)$$

where $f_a$ is the analysis frequency and $kf_a$ is the kth average-harmonic or bin frequency, $A_k(j)$ is the amplitude of the kth harmonic or bin at time frame j, $\Delta f_k(j)$ is the frequency deviation of the kth harmonic or bin at time frame j, and $\theta_k(0)$ is the initial phase of the kth harmonic.

While for pitched sounds, $f_a$ is an estimate of the fundamental frequency of the sound and $k$ corresponds to the harmonic number, for non-pitched sounds, $k$ merely corresponds to the bin number of the Fourier analysis. Since the pitched sounds were all played at Eb4, $f_a$ was selected to be 311 Hz for all of the sounds. We verified by resynthesis that our analysis model adequately described both types of sounds; the resynthesized sounds using Eq.1 were virtually indistinguishable from the originals. Some of the general methodology used here is similar to that used in McAdams *et al.* (1999).

Normalization of Rise Time

For the normalization of rise time, since no standard procedure for estimating rise times has gained widespread acceptability, we simply obtained our best visual estimates from the time-varying RMS amplitude functions (see examples in Figure 13). The log-average of these values for all sounds in the stimulus set turned out to be 22 msec. We then linearly warped the harmonic amplitude and frequency functions so that the data for the original rise-time intervals were mapped into the new average rise time interval. Also, the original remainder time interval data were mapped into the new remainder time

intervals in order to preserve the original durations. In subsequent discussions we will assume that the $A_k(j)$ functions have already been rise-time-normalized.

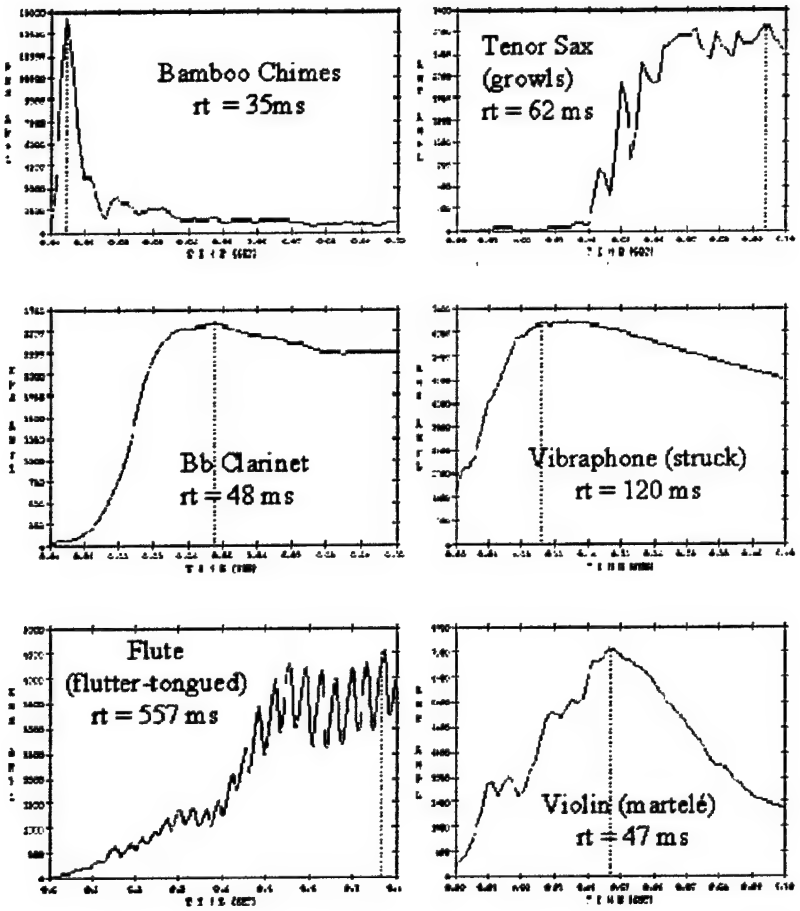



**Figure 13.** Examples of rise times estimated by the experimenters through visual inspection. The figure illustrates the difficulty of developing objective, unequivocal measures of attack or rise time.

Spectral Centroid Calculation

For each sound, the time-varying spectral centroid $f_c(j)$ was computed as an amplitude-weighted sum of all bin frequencies within each frame, multiplied by $f_a$; $f_a$ is also subtracted from this result to reduce the centroid to zero when only the first bin is active.

$$f_c(j) = \left( \frac{\sum_{k=1}^{K} k A_k(j)}{\sum_{k=1}^{K} A_k(j)} - 1 \right) f_a \qquad (2)$$

Normalization of Average Spectral Centroid

Centroid normalization proceeded in two steps. First, for each sound, we computed the RMS-amplitude-weighted average centroid over a series of J frames, under the assumption that louder portions of the tone are more important than softer portions:

$$f_{c_{aw}} = \frac{\sum_{j=0}^{J-1} A_{rms}(j) f_c(j)}{\sum_{j=0}^{J-1} A_{rms}(j)} \quad (3)$$

where the RMS amplitude for each frame was computed using:

$$A_{rms}(j) = \sqrt{\sum_{k=1}^{K} A_k^2(j)} \quad (4)$$

Computed in this manner, the log-average centroid for the entire unmodified stimulus set was found to 982 Hz. In the second step, we carried out centroid normalization by modifying the spectrum through the application of a sloping filter:

$$A_k'(j) = k^p A_k(j) \quad (5)$$

If $p$ is positive, the centroid increases, whereas if $p$ is negative, it decreases, so that centroid becomes a monotonically increasing function of $p$. By substituting Eq. 5 into Eq. 3, we arrived at the following formula for the modified centroid in terms of $p$, which was used in the iteration process:

$$f_c'(j) = \left( \frac{\sum_{k=1}^{K} k^{p+1} A_k(j)}{\sum_{k=1}^{K} k^p A_k(j)} - 1 \right) f_a \quad (6)$$

For each iteration, Eq. 6 was averaged using Eq. 3, and Newton's method was used to arrive at a value of $p$ that matched the average centroid value of 982 Hz. Finally, in order to preserve the original sound's RMS amplitude envelope, the resulting harmonic or bin amplitudes were modified as a function of frame number using

$$A_k''(j) = \frac{A_{rms}(j)}{A_{rms}'(j)} A_k'(j) \quad (7)$$

Finally the stimuli were resynthesized using these amplitudes in Eq. 1. It is important to emphasize that our technique generated an average spectral centroid value that was obtained over the duration of the signal, but did not remove local variations in centroid. Imposing the same value of centroid at each time point in the signal would have created an unnecessarily rigid constraint and would have removed much, if not, all of the timbral identity of each signal.

Participants were tested in a double-walled sound attenuation chamber (IAC), lined with Sonex textile panels to minimize acoustic reflections. Stimulus presentation and response recording was carried out with the psychoacoustic program *PsiExp* (B. Smith, 1994) running on a NeXT computer. Stimuli were reproduced on Sennheiser HD265 headphones that were connected directly to the headphone output jack on the computer. Headphone output was set to yield a comfortable listening level. On each trial, participants rated the timbral similarity of two normalized sounds presented in sequence and separated by a 2s pause. Participants entered their ratings by adjusting the position of

a sliding switch on a similarity scale whose left and right endpoints of the scale were labeled "very similar" and "very different," respectively. The scale comprised approximately 500 discrete positions on the computer monitor between these endpoints. Participants were instructed to use the full range of the similarity scale. A replay button permitted participants to hear the sound pair as many times as they wished.

Similarity judgments were obtained for the harmonic, percussive, and combined stimulus sets, presented in counterbalanced blocks of trials except that the combined set always followed the other two; we wanted participants to be familiar with both the harmonic and percussive stimuli before hearing them in a combined context. Before beginning the experimental session, each participant was given the opportunity to listen to and compare all 35 stimuli using a sound player program (SoundWorks). No specific time limit for this listening period was imposed, but participants took an average of 12 minutes before they indicated they were able to establish some basic criteria for making timbral comparisons. Participants then received 20 practice similarity judgments; none of these practice comparisons appeared in the experimental session. After the practice session, participants completed 170 trials for the harmonic and percussive sets, and 210 trials for the combined set. Trials comprised each possible pairing of sounds in randomized order. The experiment was administered across two sessions lasting approximately 1.5 hrs each.

### 4.4.2 MDS Analyses

Similarity data were analyzed with the CLASCAL model (Winsberg & De Soete, 1993). CLASCAL extends the traditional weighted Euclidean distance model (e.g., Carroll and Chang, 1970) by being able to detect a small number of latent classes or subpopulations of participants who weight the dimensions of the CLASCAL space differentially. CLASCAL also provides the option of computing a measure of variance or "specificity" that is associated uniquely with each stimulus, denoting properties of that stimulus that are not accounted for by the common dimensions of the space. Lakatos (2000) describes the general properties of this model as they pertain to the analyses of the original unmodified stimuli, and we retained the identical CLASCAL procedures for analyzing the similarity data from the current study.

### 4.4.3 Calculation of Physical Correlates

Several measures based on the time-variant spectra of the stimuli were tested for their ability to correlate with MDS-derived dimensions. The measures that we chose to investigate are *spectral centroid variation*, *spectral irregularity*, *spectral incoherence*, *spectral density*, and *decay rate*. These measures vary in applicability depending on whether the sound in question is percussive or harmonic in nature.

Spectral Centroid Variation

The amount of centroid variation was computed as the RMS deviation of the centroid compared to its amplitude-weighted average value. Amplitude averaging was used to ensure that high amplitude values of centroid received the strongest weight. Our assumption was that not only are low amplitude centroids not weighted strongly in perception but that they are also more likely to be corrupted by noise in the signal. Thus, the formula for centroid variation we arrived at is

$$\Delta f_{c_{ave}} = \left( \frac{\sum_{j=0}^{J-1} A_{rms}(f_c(j) - f_{c_{ave}})^2}{\sum_{j=0}^{J-1} A_{rms}} \right)^{\frac{1}{2}} \quad (8)$$

Finally, we used the percentage centroid variation given by

$$PCV = 100 \times \Delta f_{c_{ave}} / f_{c_{ave}} \quad (9)$$

This measure applies equally well for harmonic and percussive sounds. Values for the 36 stimuli are given in Table 1.

Spectral Irregularity:
Spectral irregularity is measured fundamentally in the frequency domain. This is a measure of the jaggedness of a spectrum. A very smooth spectrum would theoretically have zero spectral irregularity. In practice, we measure each harmonic or bin's amplitude deviation from the average of itself and its two neighbors (we necessarily exclude the first and last harmonics). For harmonic/bin $k$ the average at frame $j$ is given by

$$\overline{A}_k(j) = (A_{k-1}(j) + A_k(j) + A_{k+1}(j))/3 \quad (10)$$

Then the time-varying spectral irregularity is defined as

$$SIR(j) = \frac{\sum_{k=2}^{K-1} \overline{A}_k(j) \left| A_k(j) - \overline{A}_k(j) \right|}{\sum_{k=2}^{K-1} \overline{A}_k(j) A_{rms}(j)} \quad (11)$$

Because of the division by $A_{rms}(j)$, $SIR(j)$ is normalized to less than 1. We then amplitude-average $SIR(j)$ over time to achieve a single spectral irregularity value for the entire sound:

$$SIR_{ave} = \frac{\sum_{j=0}^{J-1} A_{rms}(j) SIR(j)}{\sum_{j=0}^{J-1} A_{rms}(j)} \quad (12)$$

Values of $SIR_{ave}$ are given for the harmonic stimuli in Table 2.

| Instrument | Centroid SD (%) | Spectral Irregularity | Spectral Incoherence | Decay T1 | Slope1 | T2 | Slope2 | T3 | Slope3 | T4 | Slope 4 | Spectral Density |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bamboo Chimes | 38.2 | | 0.6170 | 0.04 | -207.4 | 0.15 | -44.7 | 0.92 | end | | | 26.3 |
| Bongo Drum | 15.1 | | 0.4057 | 0 | -422.8 | 0.05 | -39.6 | 0.84 | end | | | 29.5 |
| Castanets | 23.8 | | 0.349 | 0 | -344.9 | 0.05 | -91.1 | 0.12 | -39.8 | 1.07 | end | 28.6 |
| Celesta | 45.5 | | 0.0870 | 0.07 | -12 | 0.74 | -18 | 1.45 | end | | | 370.0 |
| Bb Clarinet | 3.9 | 0.1775 | 0.0685 | | | | | | | | | |
| Tenor Crumhorn | 2.3 | 0.1144 | 0.0326 | | | | | | | | | |
| Cuica | 29.6 | | 0.1001 | | | | | | | | | 49.2 |
| Cymbal (bowed) | 21.3 | | 0.4628 | | | | | | | | | 78.2 |
| Cymbal (struck) | 24.8 | | 0.4853 | 0 | -17.6 | 1.45 | end | | | | | 33.0 |
| Double Bass | 65.5 | 0.1356 | 0.1357 | 0.07 | -132.5 | 0.21 | -22.5 | 0.46 | -72 | 1.04 | end | |
| English Horn | 5.8 | 0.0861 | 0.0710 | | | | | | | | | |
| Flute (flutter-tongued) | 32.3 | 0.0616 | 0.1007 | | | | | | | | | |
| Flute (no vibrato) | 6.7 | 0.1344 | 0.1182 | | | | | | | | | |
| French Horn | 14.5 | 0.0730 | 0.0416 | | | | | | | | | |
| Harp | 43.6 | 0.1608 | 0.0737 | 0.02 | -25.2 | 0.5 | -9.5 | 1.45 | end | | | |
| Harpsichord | 59.4 | 0.0600 | 0.1543 | 0.04 | -18.5 | 1.45 | end | | | | | |
| Log Drum | 84.8 | | 0.0111 | 0 | -106.1 | 0.4 | end | | | | | 35.0 |
| Marimba | 52.7 | | 0.1481 | 0.08 | -28.1 | 0.9 | -42.8 | 1.45 | end | | | 250.0 |
| Organ | 11.8 | 0.1837 | 0.2158 | | | | | | | | | |
| Piano | 15.4 | 0.0954 | 0.1336 | 0 | -21 | 0.86 | -13 | 1.45 | end | | | |
| Baroque Tenor Recorder | 12.2 | 0.1892 | 0.0177 | | | | | | | | | |
| Alto Sax | 6.4 | 0.1885 | 0.0884 | | | | | | | | | |
| Tenor Sax (growls) | 7.8 | 0.0834 | 0.1160 | | | | | | | | | 43.4 |
| Snare Drum | 37.3 | | 0.3354 | 0 | -207.5 | 0.13 | -74 | 0.65 | end | | | 54.9 |
| Steel Drum | 63.8 | | 0.6843 | 0 | -53.3 | 1.33 | end | | | | | 30.5 |
| TamTam | 55.7 | | 0.4613 | 0 | -8.6 | 1.45 | end | | | | | 33.5 |
| Tambourine (pop) | 65.9 | | 0.7493 | 0 | -.69.2 | 0.93 | end | | | | | 22.4 |
| Temple Block | 19.3 | 0.0472 | 0.0706 | 0 | -249.8 | 0.09 | -38.1 | 1.32 | end | | | 37.7 |
| Bb Trumpet (hard) | 7.6 | 0.0549 | 0.0797 | | | | | | | | | |
| C Trumpet (muted) | 17.2 | | 0.161 | | | | | | | | | |
| Tubular Bells | 1.3 | | 0.2519 | | | | | | | | | 179.5 |
| Tympani | 32.6 | | 0.2397 | 0 | -100.7 | 0.06 | -20 | 1.45 | end | | | 39.7 |
| Vibraphone (bowed) | 2.4 | | 0.0093 | 0.35 | -18.9 | 1.45 | end | | | | | 96.7 |
| Vibraphone (struck) | 2.2 | | 0.0523 | 0.1 | -19.3 | 1.45 | end | | | | | 326.7 |
| Violin (martele) | 35.1 | 0.1229 | 0.9093 | 0.05 | -145.8 | 0.47 | end | | | | | |
| Violin (no vibrato) | 5.2 | 0.1406 | 0.0722 | | | | | | | | | |

**Table 2.** The five physical correlates – spectral centroid variation, spectral irregularity, spectral incoherence, spectral density, and decay rate - of the 36 stimuli. Measures of spectral irregularity are obtained only for sounds with significant steady-state portions.

Spectral Incoherence:

Spectral incoherence is a measure of how a spectrum varies with respect to a coherent version of itself. In our case, the coherent version is based on the RMS amplitude of the signal. This version has the following properties: a) All harmonic amplitudes vary with respect to time in proportion to the RMS amplitude; i.e., the ratios between the harmonic amplitudes and the RMS amplitude are fixed. b) The ratios among the harmonic amplitudes are fixed and equal to the respective ratios of the time averages of the original harmonic amplitudes. c) The RMS amplitude of the coherent version is the same as that of the original. Note that if the original signal were completely coherent, the coherent version and the original would be the same. Now, the coherent version harmonic amplitudes are defined as

$$\hat{A}_k(j) = \frac{A_{ave_k}}{\sqrt{\sum_{k=1}^{K} A_{ave_k}^2}} A_{rms}(j) \qquad (13)$$

where the average harmonic amplitudes are given by

$$A_{ave_k} = \frac{1}{(j_{max} - j_{min})} \sum_{j=j_{min}}^{j_{max}-1} A_k(j) \qquad (14)$$

and $j_{min} = .1\,J$, $j_{max} = .9\,J$.

Finally, the single figure spectral incoherence is defined as

$$SI = \sqrt{\frac{\sum_{j=0}^{J-1}\sum_{k=1}^{K}(A_k(j) - \hat{A}_k(j))^2}{\sum_{j=0}^{J-1}\sum_{k=1}^{K}A_k^2(j)}} \qquad (15)$$

Values of spectral incoherence for the 36 stimuli are given in Table 2.

Decay Rate:
Decay rate measurements were performed on most of the percussive stimuli as well as those "harmonic" stimuli, such as the piano, which exhibit decay over their entire durations. This was done by first graphing RMS-amplitude vs. time in terms of decibels vs. time. However, it turns out that some instruments are characterized by more than one slope. Table 1 gives the beginning time and slope in dB/s for each slope for each instrument tested. The best time segments were chosen by visual estimate. Slopes were calculated using a least-squares best fit straight line for each time segment chosen. We also computed the standard deviation of fit for each segment. These data are summarized in Table 2. Note that whereas 9 stimuli only need one segment, 9 stimuli need two segments, and 2 stimuli need 3 segments.

Spectral Density:
Whereas the spectral density of harmonic sounds is obviously determined by the spacing of the harmonics and so is unity divided by the fundamental frequency, this is not so obvious for the percussion instruments where the density can vary from *low* for bar vibrators like marimba and chime to *very high* for membrane and plate instruments like snare drum and cymbal. Actually, we decided to report inverse spectral density (ISD) in order to avoid numbers that were much less than one. Therefore, we are reporting an average spacing between spectral peaks in Hz for each stimulus. To obtain these values, we first reanalyzed the percussion tones using an analysis frequency of 20 Hz, which is small enough to resolve most modes but large enough to resolve the time structure quite well. The ISD were measured at a standard time of 0.1 s after onset in order to capitalize on the richest part of the time-varying spectrum, which normally occurs at the beginning of percussion sounds. All bin numbers corresponding to amplitudes which were within 40 dB of the maximum spectral amplitude were counted and retained. These were considered to be *peaks*. The average peak-to-peak bandwidth could be computed by adding the individual bandwidths and by dividing by their number. However, this turns out to be just

$$ISD = \frac{f_{max} - f_{min}}{n-1} \qquad (16)$$

where $f_{min}$ is the minimum peak frequency, $f_{max}$ is the maximum peak frequency, and n is the number of peaks. Results for ISD for the percussion instruments are given in Table 2.

### 4.4.4 Results

BIC statistics indicated an optimal CLASCAL solution with two dimensions and with specificities, as well as two latent classes of participants (Table 3). Participants in Class 1

| | One Class | | Two Classes | |
|---|---|---|---|---|
| | **Specificites** | **No Specificites** | **Specificites** | **No Specificities** |
| Harmonic | | | | |
| 1 Dimension | -5150 | -2411 | -5216 | -2325 |
| 2 Dimensions | -5161 | -4124 | **-5388** | -4507 |
| 3 Dimensions | -5089 | -4540 | -5307 | -4826 |
| Percussive | | | | |
| 1 Dimension | -3970 | -950 | -4244 | -1210 |
| 2 Dimensions | -4140 | -3200 | -4095 | -3343 |
| 3 Dimensions | -4111 | -3995 | -4052 | **-4333** |
| 4 Dimensions | -4058 | -4108 | -3988 | -4161 |
| COMBINED | | | | |
| 1 Dimension | -6389 | -2330 | -6114 | -2384 |
| 2 Dimensions | -6475 | -5288 | -6386 | -5469 |
| 3 Dimensions | **-6558** | -6182 | -6463 | -6287 |
| 4 Dimensions | -6514 | -6052 | | |

**Table 3.** Values of information criterion BIC for CLASCAL models derived from similarity ratings in comparisons of harmonic, percussive, and combined stimulus sets. The model with the lowest value for the BIC criterion (in boldface) is considered to be the most appropriate.

weighted dimension 1 more heavily than did those in Class 2. For these individuals, the categorical distinction between the steady-state brass and woodwind sounds and the impulsive string sounds seemed to be a primary one. Those in Class 2 weighted dimension 2 and specificities more heavily, indicating that these individuals placed greater emphasis on the unique properties of a subset of sounds.

Figure 14 displays the two-dimensional CLASCAL spaces for the original harmonic data set from Lakatos (2000), representing the ratings of professional musicians and nonmusicians, and the set equalized for spectral centroid and rise time in the current study, representing a separate group of nonmusicians only. Although caution should be exercised in drawing comparisons across these two solutions, especially because of the different samples of participants, we find it informative to highlight some of the general similarities in the structures of the solutions that remain despite the spectral and temporal modifications to the 18 original stimuli. Drawing comparisons across participants with differing levels of musical training seems somewhat justified since Lakatos (2000) found no significant differences in the weightings of the CLASCAL dimensions for the unmodified stimulus sets as a function of musical expertise. Timbres represented by larger filled circles have specificities greater than 1.0, or greater than 10% of their variance that is not accounted for by the common dimensions of the CLASCAL space.

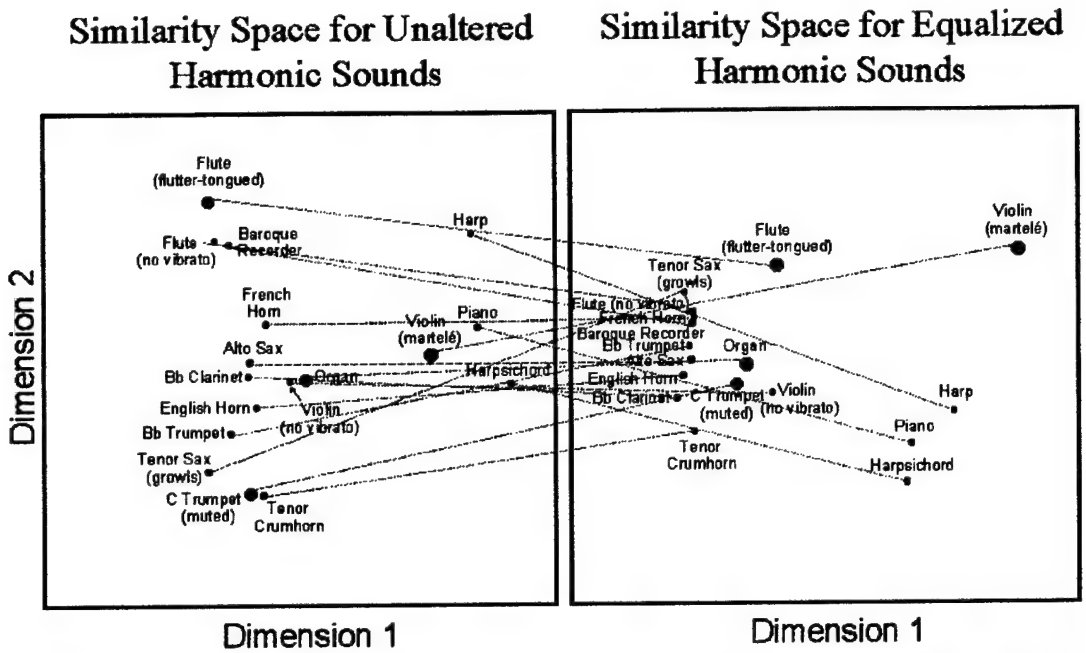



**Figure 14.** Two-dimensional CLASCAL spaces for: (a) the original harmonic stimulus set from Lakatos (2000), and (b) the modified set tested in the present study. Dashed lines show the changes in position of individual timbre across the two solutions. Large circles in both figures denote timbres with specificities greater than 1.0.

The primary spatial feature preserved across original and modified stimulus sets is the segregation of timbres with significant steady-state, sustained segments separated from those with transient characteristics along dimension 1. These latter "percussive" harmonic stimuli are high on dimension 1, with rapid decay (violin martelé) being higher and slow decays (harp, piano, and harpsichord) have lower values on both dimension 1 and 2. The sustained tones are clustered in a region of in the low range of dimension 1 and medium range of dimension 2. There does not appear to be any particular ordering within this sustained group for the modified stimulus set, in contrast to the original space, where spectral centroid and its frequent perceptual correlate, "brightness," clearly differentiated sounds along dimension 2. The tenor saxophone and flutter-tongued flute, both of which can be categorized as "rough," are close together in space. Interestingly, several of the instruments that had high specificity values in the original study (the organ, flutter-tongued flute, organ, and muted C trumpet) also have high specificities in the modified space; apparently, these stimuli continue possess features that are not well represented by the common dimensions of the space. Some of these sounds - the flutter-tongued flute and the tenor sax growls, for instance - have noise components or other inharmonic properties, while others, such as the muted trumpet or the violin martelé, have unusual modes of excitation or distinctive source features relative to the other stimuli. Correlations between the dimensions of the spaces for the original and modified sets are shown in Table 4. There is strong correlation (.89) between the first dimensions of the two spaces, suggesting that equalization of attack time did not substantially alter the bimodal distribution of stimuli along this dimension. There was little correlation between the second dimensions of the spaces (.32), however, revealing that the equalization of centroid shifts the similarities around for these sustained pitched instruments. In particular, the sustained woodwind and brass sounds appear to collapse along dimension 2 relative to the impulsive string sounds across original and modified harmonic sets.

| | Dimension 1 | Dimension 2 | Dimension 3 |
|---|---|---|---|
| **Harmonic.** | .89 | .32 | |
| **Percussive** | .75 | .79 | .66 |
| **Combined** | .91 | .57 | |

**Table 4.** Correlations between the dimensional coordinates of the original stimuli used in Lakatos (2000) and the modified stimuli tested here. Strong intercorrelations for most dimensions suggests that much of the perceptual structure of the CLASCAL spaces is preserved even when stimuli are equalized for spectral centroid and rise time.

Table 5 displays the correlations between the dimensions of the CLASCAL solutions for the three stimulus sets and the five acoustic measures described above: Centroid variation, spectral irregularity, spectral incoherence, decay, and spectral density.

| | Centroid SD (%) | Spectral Irregularity | Spectral Incoherence | Decay | Spectral Density |
|---|---|---|---|---|---|
| **Harmonic** | | | | | |
| **Dimension 1** | .778 | .458 | .626 | --- | --- |
| **Dimension 2** | -.142 | -.169 | .412 | --- | --- |
| **Percussive** | | | | | |
| **Dimension 1** | .024 | -.468 | .363 | -.817[a] | -.479 |
| **Dimension 2** | -.007 | .283 | -.568 | .084[a] | .504 |
| **Dimension 3** | .092 | -.569 | .553 | .093[a] | -.268 |
| **Combined** | | | | | |
| **Dimension 1** | -.656 | .067 | -.575 | .675[b] | .268[c] |
| **Dimension 2** | .154 | .210 | -.341 | .822[b] | .698[c] |
| **Dimension 3** | .019 | -.148 | -.380 | .220[b] | .255[c] |

[a]percussive sounds with significant decays (14 sounds)
[b]combined sounds with significant decays (11 sounds)
[c]combined sounds with inharmonic spectra (percussive sounds)

**Table 5.** Correlations (r) between stimulus dimensional coordinates and stimulus values for five acoustical attributes: centroid standard deviation (expressed as percentages), spectral irregularity, spectral incoherence, decay, and spectral density (inverse values for density). Decays for harmonic stimulus set are not shown since decay accounts for the bimodal distribution of stimuli along Dimension 1. Density measures are constant for sounds with harmonically related partials and are therefore omitted for the harmonic set.

Although multiple, stepwise regression analyses would seem appropriate for analyzing the relative contributions of the acoustic measures to each dimension, we resisted doing so at this point because it is unclear to us at this point whether these measures are perceptually, as well as physically orthogonal, as linear regression models assume; in

order to avoid such a strong assumption, we simple list the correlations separately between each acoustic measure and MDS dimension. We did not calculate spectral density measures for the harmonic set since such calculations for sounds with harmonically, or quasi-harmonically, related partials would yield near identical values. In a similar vein, we did not compute decay measures because we imposed artificial 50 ms linear decay ramps on the steady-state portions of the woodwind and brass instruments in order to bring their duration to 2000 ms, the standard duration for all stimuli. It is nonetheless clear from Figure 1b that there is a clear separation along dimension 1 between steady-state and impulsive sounds and, given that rise times were equalized, this separation must be based on the prominent decay values for the latter groups of signals.

Stimulus coordinates along dimension 1 correlate moderately-to-strongly with the standard deviation of the centroid ($r = .778$), spectral irregularity ($r = .458$), and spectral incoherence ($r = .626$). In other words, stimuli with high values along dimension 1 have high values for these three acoustic correlates. An ecological interpretation may best explain these multiple correlations: a string that is excited impulsively by a hammer, bow, or finger, for example, and that is left to decay will generate signal that, over time, has a progressively lower centroid, and less spectral irregularity/incoherence because the higher-frequency partials die out more quickly than do the lower-frequency ones. Instruments excited continuously will have significant steady-state segments that show far less fluctuation in these regards. Dimension 2 may be interpreted more easily in purely spectral terms: instruments with greater spectral incoherence, such as the flute and the tenor saxophone – tend to be located higher on this dimension. This correlation ($r = .412$) is far from perfect, however, with some instruments, such as the muted trumpet, located low on dimension 2 despite relatively high spectral incoherence.

In sum, the CLASCAL solution for the harmonic stimuli equalized for centroid and rise time preserves much of its bimodal distribution along dimension 1, indicating that signal decay, as well as rise time, are important correlates. Signal decay cannot be considered independently of changes in the signal spectrum, however, as the substantial correlations in Table 5 indicate. Removal of spectral centroid as a differentiating factor has a more profound effect on the organization of stimuli along dimension 2. According to participants' reports and our own informal listening, the steady-state sounds are more strongly affected by this equalization process, with the woodwind instruments in particular losing much of their distinctiveness. Nonetheless, even along dimension 2 many of the stimuli maintain their relative location, albeit in a compressed arrangement, when compared to the CLASCAL solution for the original harmonic stimuli.

Percussive Set

The optimal CLASCAL solution for the percussive stimuli turned out to be three dimensional without specificities, just as for the original, unmodified set (see Table 2). CLASCAL detected two classes of participants; Class 2 weighted the three dimensions more heavily than did Class 1, especially for dimension 3. Correlations between the three dimensions of the spaces for the original and modified stimulus sets are all moderately strong, ranging from .66 to .79 (see Table 3), a somewhat surprising finding given the equalization process the stimuli in the current study were subjected to.

Figure 15 comparisons dimensions 1 and 2 for the original and modified percussive stimulus sets. Four general clusters of instruments are apparent in Figure 15: (1) The

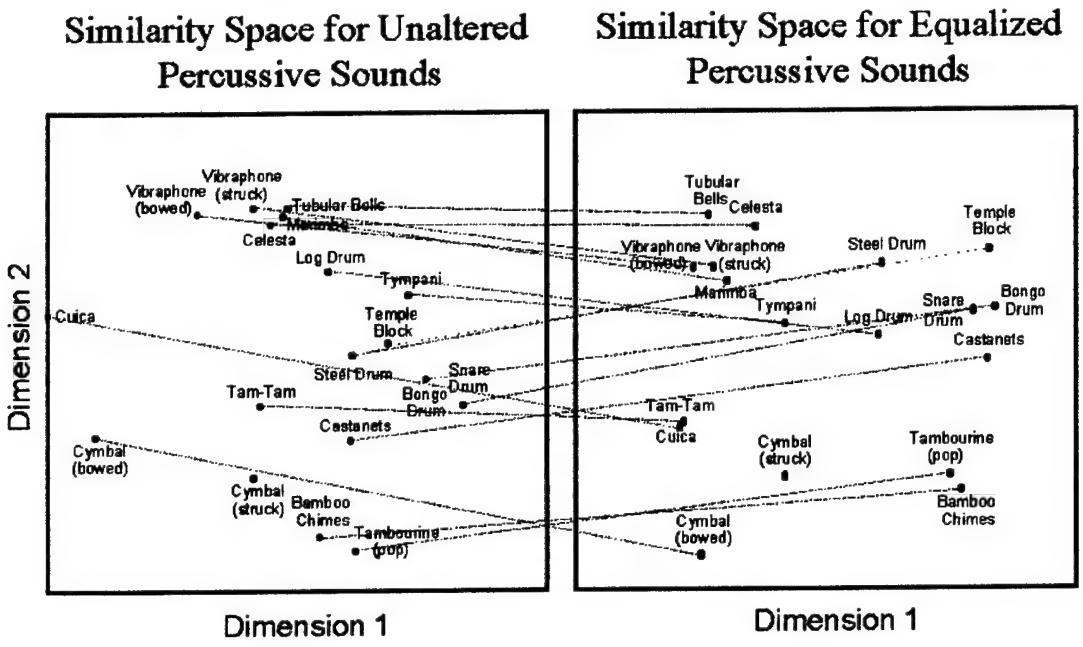



**Figure 15.** Dimensions 1 and 2 of the three-dimensional CLASCAL space for percussive stimuli are shown for the original percussive set from Lakatos (2000) and the modified set used in the present study.

tubular bells, celesta, marimba, and bowed and struck vibraphones, which are highly pitched with fairly long decays; (2) The steel drum, temple block, log drum, snare drum, bongo drum, and castanets, which are weakly pitched and have relatively fast decays (the tympani may be a transitional instrument between these first two groups); (3) The tam-tam cuica, and struck/bowed cymbals have long decay times; (4) The tambourine and bamboo chimes are both temporally short and have granular spectral structures. If we compare the positions of the instruments across the original and modified spaces, some locations do change: For example, the tam-tam and cuica, which are relatively far apart in the original study are right next to one another in the modified study (they separate along dimension 3). But the spaces have numerous general similarities: most instruments end up in roughly the same quadrants across the two spaces.

Signal decay is a strong correlate of dimension 1 ($r = -.817$), with rapidly decaying sounds, primarily the drums, high along the dimension and more slowly decaying sounds, such as the bowed bars, plate, and membranes, on the low side. Additionally, spectral irregularity and spectral density are moderately correlated with dimension 1. Since the number of partials typically decreases as the sound decays and the signal spectrum also tends to become less jagged over this period, it is not unexpected that these two acoustic features also correlated with this dimension. Spectral density ($r = .504$) and spectral incoherence ($r = -.568$) account for a substantial portion of the variance along dimension 2. Instruments with high spectral density, particularly ones with metallic structures, such as the tam-tam, tambourine, and cymbal, are high on dimension 2. The bamboo chimes, although not metallic, consist of multiple components and also generate a rich spectral structure when struck in sequence.

Dimensions 1 and 3 are compared for the original and modified stimulus sets in Figure 16. It might appear that spectral density is a good correlate of dimension 3. However, the steel drum, tambourine, bamboo chimes, bongo drum, temple block, and castanets, which

form a large cluster, do not appear to follow that rule. Rather, spectral irregularity (r = -.569) and incoherence (r = .553) emerge as moderately strong correlates of dimension 3. As in Figure 2b, the tubular bells, bowed and struck vibraphones, bowed cymbal, celesta and marimba seem to form a cluster of pitched instruments with medium decays.

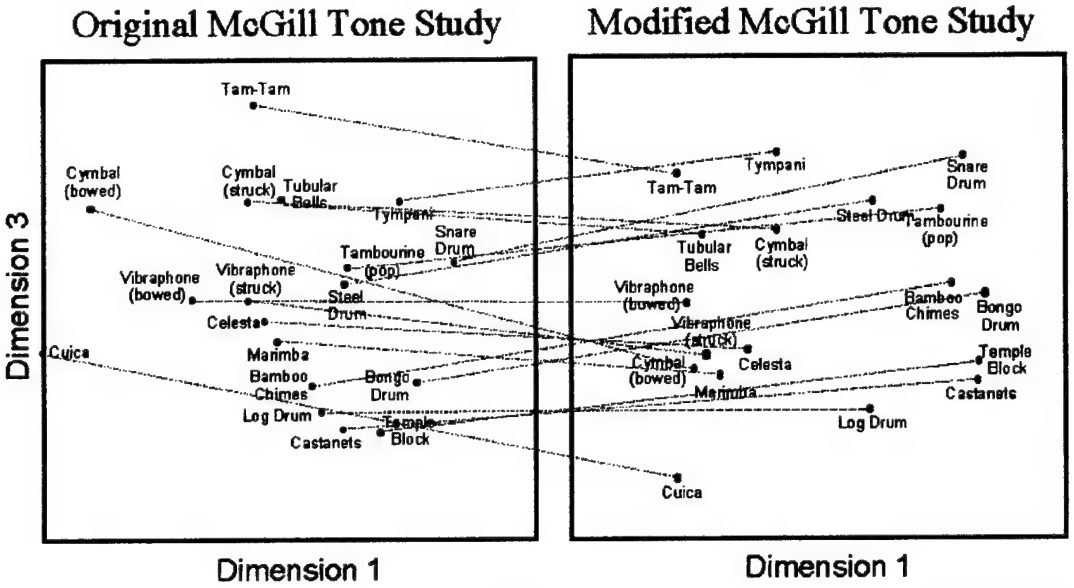



**Figure 16.** Dimensions 1 and 3 of the CLASCAL space for percussive stimuli are shown for the original percussive set from Lakatos (2000) and the modified set used in the present study.

Combined Set

CLASCAL chose a three-dimensional solution with specificities for the combined stimulus set (see Table 3). Figure 17 shows the optimal CLASCAL solutions for

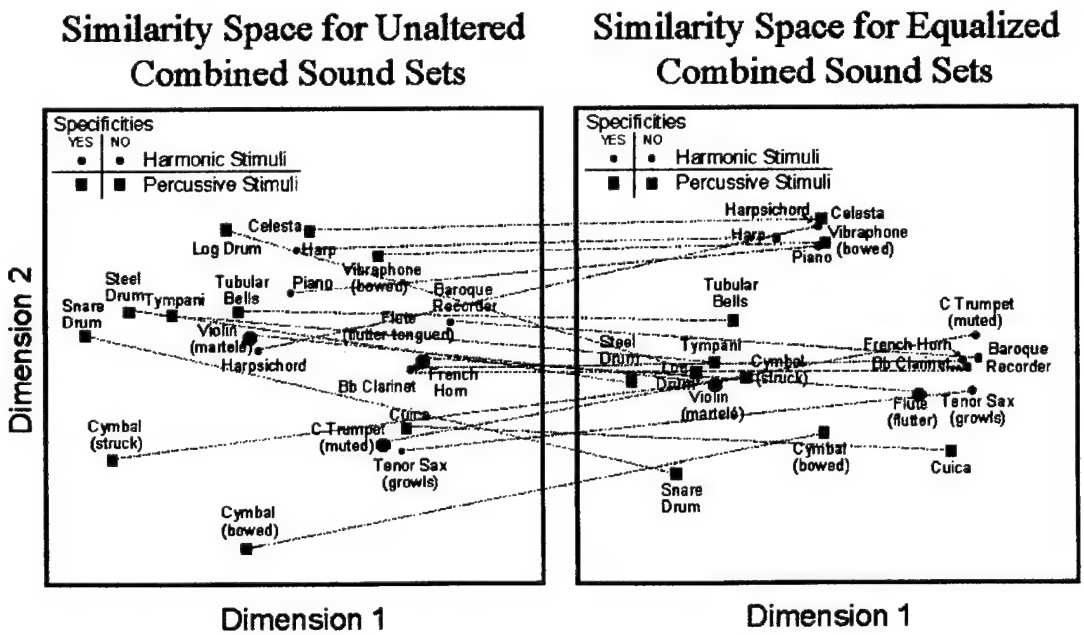



**Figure 17.** Dimensions 1 and 2 for the CLASCAL combined space are compared for the original combined set from Lakatos (2000) and the equivalent modified set. Harmonic stimuli are indicated with circles, percussive stimuli with squares. Shaded icons denote timbres with specificities > 1.0.

dimensions 1 and 2 of the original and modified combined stimulus sets. Stimuli from the harmonic set are represented by circles while percussive stimuli are indicated by squares; harmonic and percussive stimuli with specificities greater than 1.0 are indicated by gray shading. There appear to be three main clusters of instruments: (1) The harpsichord, celesta, harp, bowed vibraphone, and piano are highly pitched with long decays; (2) The steel drum, snare drum, tympani, struck cymbal, log drum, violin martelé, and tubular bells are clustered as having percussive envelopes, although pitchedness does not appear to be a factor here; (3) The muted trumpet, French horn, clarinet, baroque recorder, flute, tenor saxophone, and cuica form a cluster of pitched, sustained sounds. The snare drum and bowed cymbal appear to be the only instruments do not seem to fit into any of these groups. Clusters 1 and 3 are more tightly grouped for the modified stimuli along dimension 2, as they were for the modified harmonic stimuli, reflecting the significance of the removal of spectral centroid variations from these pitched signals.

The original and modified combined stimulus sets have several high-specificity sounds in common: The flutter-tongued flute, bowed cymbal, cuica, violin (martele), and tympani. The first four of these have unusual modes of excitation: The flutter-tongued flute has a rough texture with significant noise components and the bowed cymbal is an instrument that is typically struck in Western music. Similarly, most Western membranophones are struck, not bowed like the cuica. The tympani may seem exceptional to listeners because its apparent pitch is one octave below those of the other pitched instruments tested here.
As was the case for the harmonic space, centroid variation and spectral incoherence correlated strongly, albeit negatively, with dimension 1 of the combined space, but, unlike the harmonic space, spectral irregularity did not. Rather, decay revealed itself to be a substantial correlate of dimension 1 for the percussive sounds, with sounds possessing shorter decays located at the left end of the dimension. From an acoustic standpoint, it is surprising short-decay sounds such as the snare drum and struck cymbal would also have greater fluctuations in the moment-to-moment amplitudes of its spectrum and therefore would have larger spectral centroid perturbations. Dimension 3 turned out to be more difficult to interpret acoustically (see Figure 18). Aside from a weak-to-moderate

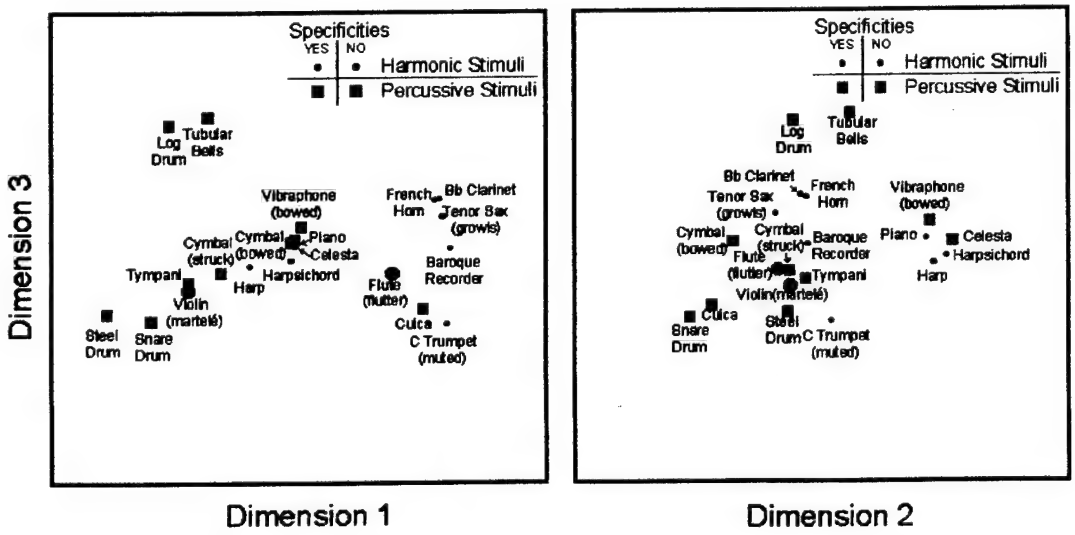



**Figure 18.** Dimension 3 plotted against dimensions 1 and 2 for CLASCAL combined space Harmonic stimuli are indicated with circles, percussive stimuli with squares. Shaded icons denote timbres with specificities > 1.0.

correlation with spectral incoherence, the remaining acoustic correlates did not account for any substantial variance along this dimension. Some participants noted after the experiment that sounds that ended up high on dimension 3, such as the log drum and tubular bells, have a "fuller, rounder" sound than sounds with lower values, such as the snare drum and cuica, which sounded "harsher" and "more strident." Overall, however, participants' *post hoc* comments were not consistent enough to yield an unequivocal interpretation for this dimension, even a verbal or semantic one. It may be that there is an acoustical feature that covaries with sounds along this dimension that simply has not been tapped by any of the measures we were able to develop for this paper.

In sum, spectral centroid and rise time may represent fundamental acoustical correlates of timbre, but the results obtained in the current study show that there are additional spectral and temporal features available to listeners when the effects of these two primary correlates are reduced by equalization techniques. The fact that the timbres of many of the stimuli used here are not destroyed by the equalization process but rather that they preserve substantial aspects of their original identity suggests that an accurate psychophysical definition of timbre should move beyond a two-factor interpretation. One might argue that there are alternative ways of modeling the central tendencies of spectra and signal rise times and that such alternative definitions might result in greater perceptual "degeneration" of the instrumental timbres tested here. For instance, although we equated sounds for a spectral centroid value averaged over the time course of each signal, we did not remove variations in centroid within the signal; undoubtedly, removing all centroid variations would radically alter the timbres of most stimuli tested here. Such potential caveats should not nullify the general implications of the current findings however, since past MDS studies that have found centroid and rise time to be significant acoustical correlates have used similar metrics for operationalizing them.

Most dimensions of the CLASCAL solutions for traditional pitched orchestral instruments, percussive instruments, and their combination, had multiple acoustic correlates. Decay, in particular, did not appear as the sole correlate for any individual dimension. For example, spectral irregularity and the standard deviation of the centroid appeared with decay as strong correlates of dimension 1 of the combined stimulus set, whereas for dimension 2, both decay and spectral density are strong correlates. Similarly, spectral irregularity and density covaried with decay for dimension 1 of the percussive CLASCAL space. The lack of orthogonal relations among these acoustic features suggests that it may be more fruitful to interpret such covariations within an ecological framework. As the amplitudes of the various excitation modes of a sound source decay, for instance, higher frequency modes tend to decay more rapidly, often leading to reductions in spectral density and spectral irregularity, as well as fluctuations in centroid. Percussive instruments fabricated from metal, such as the tam-tam and cymbals, typically have denser spectra and longer decays than non-metallic ones so that these two acoustic features will tend to covary in such instruments. In a similar vein, dense spectra often, but not always, tend to have more amplitude irregularities over the signal's duration. In sum, the covarying features we have observed here may serve as markers for certain physical properties that the instruments generating the signals share.

One unresolved issue concerns the salience of cues such as spectral irregularity and density when information about spectral centroid and rise time remains intact. Can and

do such cues ever serve as primary criteria for judging timbre in unaltered signals, or do they come to the fore only when information about centroid and attack are suppressed? If the former characterization is more accurate, why have most of these additional cues not emerged independently in past MDS research? We suspect the reason is that a frequent interrelatedness among these cues in certain classes of sound sources precludes their emergence as independent factors or dimensions in MDS analyses. Variance associated with such cues may be "absorbed" by the primary dimensions of centroid and rise time, since the cues can also covary with these dimensions, and/or they may represented in an aggregate manner by an additional, difficult-to-interpret dimension. In this sense, a purely dimensional interpretation of timbre may not suffice for more than two or three separable dimensions, and a source model of timbre may better account for the co-occurrence of additional timbral cues.

## 4.5 Interactions Among the Dimensions of Timbre

**Summary:** Although MDS studies have contributed to a descriptive model of timbre, little is yet known about how the auditory system analyzes timbral dimensions. For example, to what extent are such dimensions separable or interacting? If listeners can attend selectively to one timbral dimension while ignoring orthogonal variation on an irrelevant dimension, the dimensions can be considered separable or perceptually independent. On the other hand, if two dimensions interact, a failure of selective attention known as Garner interference occurs (Garner, 1974). In one of the few studies in this domain, Melara and Marks (1990) used a Garnerian speeded classification task to determine whether timbre was perceptually separable from pitch and loudness, and found clear interactions between these dimensions; however, the authors defined timbre as the duty cycle of a rectangular pulse, a definition that does not adequately characterize the spectral or temporal correlates of timbral dimensions reported in MDS studies. Therefore, we decided to employ the paradigm of Melara and Marks to examine possible interactions among three timbral dimensions - "brightness" (spectral centroid), "attack" (rise time), and "hollowness" (spectral flux). Given that a principal function of timbre is to convey convergent information about the physical source generating the sound, we expected to find considerable interactions between the two timbral dimensions reported consistently in the MDS literature - brightness and attack. We were less confident in our predictions for hollowness, since spectral flux may be an overly simplified approximation of what are often highly complex and varied spectral irregularities. If the extraction of such irregularities proceeds at relatively high levels of auditory processing, such processing may be functionally separate from that of centroid and attack. Therefore, we tentatively hypothesized that hollowness would be separable from brightness. In Experiment 1, participants classified values from one dimension (brightness or sharpness) while the other dimension was either held constant (baseline), correlated, or varied orthogonally (filtering). In Experiment 2, hollowness replaced sharpness. Classification for filtering tasks was poor relative to baseline for brightness and attack, but not hollowness. No redundancy gains and losses occurred for correlated dimensions in either experiment, but baseline and filtering tasks revealed congruity effects for brightness and attack. The results point to crosstalk between channels conveying information about centroid and rise time, but independent processing of spectral flux

### 4.5.1 Experiment 1

Twenty students (16 men and 4 women) were recruited and paid $8 per hour for their participation. Stimuli were complex signals with a fundamental frequency of 300 Hz, each signal comprising 30 harmonics. Signals were generated digitally by a TDT (Tucker-Davis Technologies) signal processing system operating in conjunction with a Midwest Micro PC. After digital-to-analog conversion, signals were passed through an anti-aliasing filter (8 kHz cutoff) and presented over Realistic Nova-40 headphones at 75 dB SPL. Four stimuli, each 150 ms long, were generated by selecting two values along each of two subjective dimensions of timbre: (1) "brightness," which correlates strongly with the spectral center of gravity, or centroid, of the signal; (2) "attack," whose primary physical correlate is the signal's rise time.

Figure 19a-b illustrates the nature of values selected along each of these dimensions. Since values were adjusted for each participant during a separate practice session in order

to achieve baseline classification performance of about 85%, the values presented in Figure 1 represent average settings across the ten participants. The signal to be classified as "more bright" had a spectral centroid - calculated using a measure described in Beauchamp & Horner (1995). Similarly, the signal to be classified as having a "faster attack" possessed a rise time 20 ms faster, on average, than one having the "slower attack" (30 ms vs. 50 ms, measured from start of signal to the time of its maximum amplitude).

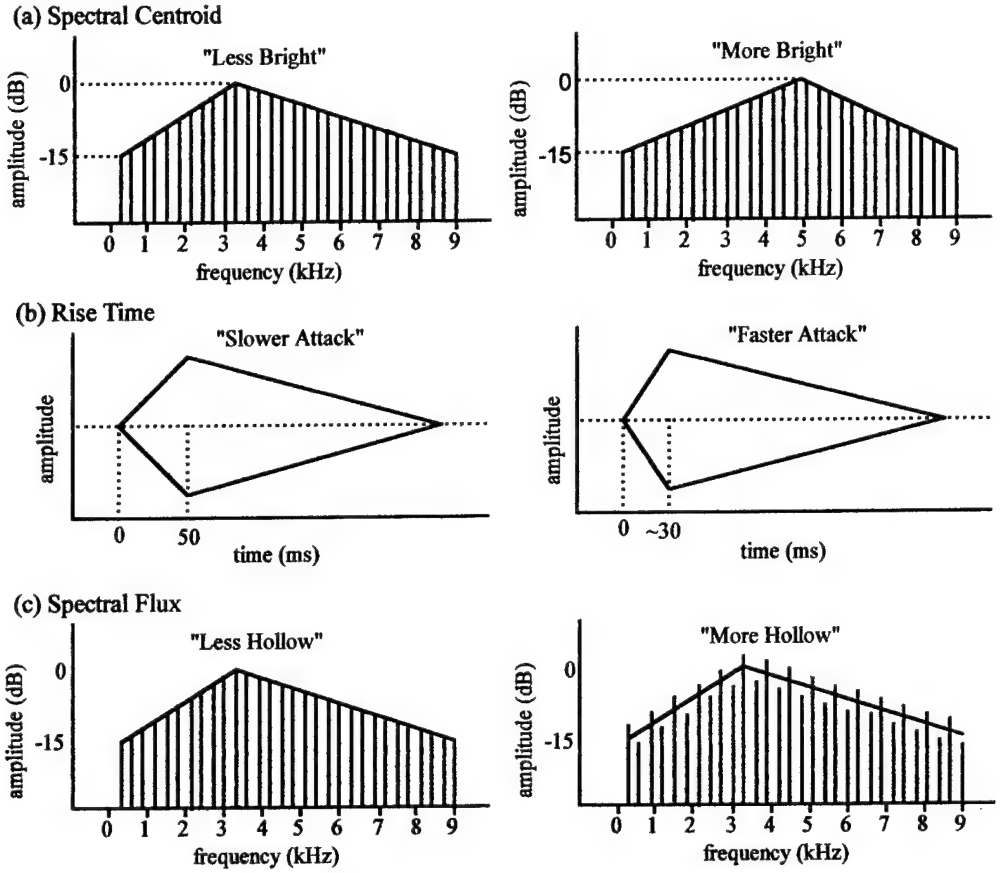



**Figure 19.** Spectral and temporal characteristics of the six signals used in Experiments 1 and 2. a) values of spectral centroid (xxx Hz and ~xxx Hz) used to generate perceptions of greater or lesser "brightness"; (b) rise times( 50 ms and ~35 ms) representing "slower" and "faster" attack; (c) standard deviations of harmonics from spectral envelope (0 dB and ~2.5 dB) yielding perception of greater or lesser "hollowness." Signals on the left side were fixed in their parameters while signals on the right side were calibrated for each subject to yield individual discrimination scores of 80-85%along each dimension.

Our experimental tasks followed closely those used by Melara and Marks (1990). Each participant completed 10 experimental tasks, each task comprising 96 trials; 5 involved classification of brightness and 5 involved classification of attack. Tasks were presented randomly during experimental sessions. Four tasks were baseline, where participants discriminated between two values along one dimension (e.g., less vs. more bright) while the other dimension was held constant at one of its two values (e.g., less bright). Four tasks involved correlated dimensions: In two of these tasks, participants discriminated stimuli whose values came from corresponding poles of their respective dimensions, while in the remaining two, they discriminated two negatively correlated, or incongruent,

stimuli, according to either brightness or attack. Our default assumption regarding congruent values in brightness and attack - to some extent an arbitrary one, was that greater brightness correlated positively with faster attack. Finally, in the two filtering tasks, participants classified all four stimuli along either the brightness or attack dimension.

Sessions were conducted in an audiometric test chamber. Participants first completed a practice session consisting of the two baseline tasks (48 trials each). Based on correct response rates in the practice session, differences in the values at one or both of the dimensions were either increased or reduced, if necessary, to bring rates to about 85% correct. A second practice session was then given to determine whether the change in values along the appropriate dimension(s) brought performance into the above range. The main experimental tasks were presented in random order, each consisting of 96 trials. Participants initiated each trial by pressing the space bar on a keyboard. Stimuli appeared 1000 ms later, separated by 500 ms. Discriminations were made by pressing one of two keys on the keyboard; key assignments were counterbalanced across participants. Participants were instructed to respond as fast as possible without sacrificing error. Participants' response speed and accuracy were displayed at the end of each trial. The experimental session lasted approximately 1-1/2 h.

Table 6 shows mean reaction times and error rates for each of the experimental tasks. Baseline discriminabilities were unequal, with participants discriminating brightness 27 ms faster on average than attack [363 ms vs 390 ms; $F(1,19)=4.52$, $p=.05$]. Pooled across brightness and attack, filtering tasks resulting in average reaction times that were 101 ms slower than those for baseline tasks [$F(1,19)=14.41$, $p=.001$]. This Garner interference was significant when tested separately for brightness [$F(1,19)=13.02$, $p=.002$] and attack [$F(1,19)=7.07$, $p=.02$]. There was no interaction in this case between task (baseline, filtering) and dimension [brightness, attack; $F(1,19)=.988$, $p=.33$]

| Task | Brightness (Centroid) | | Attack (Rise Time) | | Mean | |
|---|---|---|---|---|---|---|
| | M | Prop. Error | M | Prop. Error | M | Prop. Error |
| Baseline | 363 | .157 | 390 | .173 | 377 | .165 |
| Filtering | 484 | .316 | 473 | .322 | 478 | .319 |
| Positively Correlated Dimensions | 385 | .166 | 358 | .147 | 369 | .145 |
| Negatively Correlated Dimensions | 380 | .183 | 381 | .149 | 380 | .175 |
| Mean | 403 | .206 | 401 | .198 | 402 | .201 |

**Table 6.** Mean reaction times (in milliseconds) from Experiment 1 for brightness and attack classifications for baseline, filtering and correlated dimensions.

Contrary to our expectations, correlated dimensions were not classified significantly faster (for positive correlations) or slower (for negative correlations) than baseline. Pooled across brightness and attack tasks, participants classified positively correlated dimensions in 8 msec faster than baseline [369 ms vs. 377 ms; $F(1,19)=.49$, $p=.49$]. Negatively correlated dimensions were classified 6 msec slower [383 ms. vs 377 ms. $F(1,19)=.08$, $p=.78$]. Therefore, no redundancy gains or losses occurred for brightness and attack.

Table 7 displays congruity scores computed separately for baseline, filtering and correlated tasks. Across all tasks, congruent stimuli were classified 38 ms faster than incongruent stimuli [F(1,19)=8.54, p=.009]. Congruence effects remain significant then reaction times are pooled separately for centroid F(1,19)=4.85, p=.04] and attack [F(1,19)=6.83, p=.02]. A congruence effect was also observed in the baseline tasks alone [F(1,19)= 12.77, p=.002] and in filtering tasks [F(1,19)=3.70, p=.05]; only for correlated tasks, as noted above, was the effect non-significant [F(1,19)=1.62, p=.22].

| | Brightnness (Centroid) | | | Sharpness (Rise Time) | | |
|---|---|---|---|---|---|---|
| | Stimuli | | Congruity | Stimuli | | Congruity |
| Task | Congruent | Incongruent | Score | Congruent | Incongruent | Score |
| Baseline | 325 | 401 | +76 | 366 | 413 | +47 |
| Filtering | 473 | 495 | +22 | 444 | 502 | +58 |
| Correlated | 381 | 380 | -1 | 358 | 385 | +27 |
| Mean | 393 | 425 | +32 | 389 | 433 | +44 |

**Table 7.** Mean reaction time (in milliseconds) from Experiment 1 for stimuli with congruent and incongruent attributes, shown separately for brightness and attack and for baseline, filtering and correlated dimensions.

### 4.5.2 Experiment 2

The twenty participants from Experiment 1 also completed Experiment 2. Half of the participants began with Experiment 2 first; the remaining half first completed Experiment 1. Stimuli were again complex signals comprising 30 harmonics. As before, four stimuli were generated by selecting two values along each of two subjective dimensions: (1) "brightness," with parameters retained from Experiment 1; (2) "hollowness," which we predicted would correlate positively with the standard deviation, or spectral flux, of the harmonic amplitudes. Figure 19c shows how spectral flux varied across the two values selected for Experiment 2. The harmonics of the signal to be classified as "less hollow" conformed closely to the spectral envelope outlined in Figure 1c, while the harmonics of the signal to be classified as "more hollow" deviated from this envelope by an average of approximately 2.5 dB (the amount of deviation was calibrated separately for each subject to yield discrimination scores of 80-85%, but all deviations fell in the range of 1-4 dB). Rise time was held constant at 50 ms for all four stimuli. Aside from substituting "hollowness" for "attack," the tasks, procedures, and stimulus generation methods from Experiment 1 were retained.

Mean RT's and error rates for the tasks of Experiment 2 are shown in Table 8. Across baseline tasks, RT's for brightness and hollowness did not differ significantly [F(1,19)=.016, p=.90]. The average response time for filtering tasks was slower than for baseline [F(1,19)=15.13, p=.001], but for individual dimensions, the effect is significantly only for brightness [F(1,19)=12.90, p=.002], not hollowness [F(1,19)=3.54, p=.08]. No interaction occurred between task (baseline, filtering) and dimension [brightness, attack; F(1,19)=.751, p=.40]. Surprisingly, no congruity effects occurred for positively [F(1,19)=3.23, p=.09] or negatively [F(1,19)=.02, p=.89] correlated dimensions.

| Task | Brightness (Centroid) | | Hollowness (Flux) | | Mean | |
|---|---|---|---|---|---|---|
| | M | Prop. Error | M | Prop. Error | M | Prop. Error |
| Baseline | 380 | .200 | 375 | .176 | 378 | .188 |
| Filtering | 447 | .218 | 417 | .233 | 432 | .225 |
| Positively Correlated Dimensions | 340 | .114 | 348 | .118 | 343 | .116 |
| Negatively Correlated Dimensions | 374 | .131 | 383 | .145 | 378 | .118 |
| Mean | 385 | .166 | 381 | .168 | 383 | .169 |

**Table 8.** Mean reaction time (in milliseconds) from Experiment 2 for stimuli with congruent and incongruent attributes, shown separately for brightness and attack and for baseline, filtering and correlated dimensions.

Congruity scores are shown in Table 9 for baseline, filtering and correlated tasks. Pooling across all tasks, the congruence effect just miss significance [376 ms vs. 400 ms. $F(1,19)=4.02$, $p=.06$]. For separate tasks, congruence effects are significant for baseline tasks [$F(1,19)=5.26$, $p=.03$], but not for either correlated [$F(1,19)=2.36$, $p=.14$] or filtering tasks [$F(1,19)=.12$, $p=.74$]. The lack of significant congruity effects, in spite of seemingly large congruity scores in for several tasks in Table 4, can be attributed in part to large individual differences in RTs among subjects, particularly for hollowness. This can be seen when congruity scores are calculated separately for brightness and hollowness: The effect is significant for brightness [$F(1,19)=4.14$, $p=.05$], as in Experiment 1, but not for hollowness [$F(1,19)=3.05$, $p=.59$].

| Task | Brightnness (Centroid) | | | Hollowness (Spectral Flux) | | |
|---|---|---|---|---|---|---|
| | Stimuli | | Congruity Score | Stimuli | | Congruity Score |
| | Congruent | Incongruent | | Congruent | Incongruent | |
| Baseline | 331 | 401 | +70 | 382 | 369 | -13 |
| Filtering | 463 | 430 | -33 | 390 | 443 | +53 |
| Correlated | 340 | 374 | +34 | 348 | 383 | +35 |
| Mean | 378 | 402 | +24 | 373 | 398 | +25 |

**Table 9.** Mean reaction time (in milliseconds) from Experiment 2 for stimuli with congruent and incongruent attributes, shown separately for brightness and attack and for baseline, filtering and correlated dimensions.

In sum, Experiment 1 revealed an interaction between brightness and attack. In filtering tasks, participants showed large interference effects relative to baseline; that is, they were unable to attend selectively to either dimension in the face of variation along the irrelevant dimension. In addition, congruence effects occurred for baseline and filtering tasks, but no significant redundancy gains or losses were obtained for positively or negatively correlated dimensions. The results indicate that brightness and attack interact integrally, but do not possess a polar correspondence of attributes. Experiment 2, on the other hand, suggests that hollowness is perceptually separable from brightness. No significant interference effects were observed hollowness in filtering tasks, nor did this dimension show significant congruence effects.

Taken as a whole, the two experiments complement past MDS studies of timbre by shedding additional light on its dimensional nature. Evidence that brightness and attack - the most common perceptual correlates of timbral dimensions - interact integrally points to considerable cross-talk between these dimensions, possibly at an early

sensory/perceptual level. The perceptual independence of hollowness suggests that auditory analysis of spectral flux may proceed independently from the encoding of spectral central tendency and rise time, perhaps at higher processing levels. We suspect that spectral flux, as defined in this study and by Krimphoff et al. (1994), may be only a prototypical case that subsumes a wide range of spectral irregularities found in environmental and artificially generated sounds. If this is the case, it may not surprising, then, that differing interpretations of this dimension have been reported in MDS studies that report three or more dimensions of timbre.

## 4.6 Loudness-Independent Cues to Object Striking Force

**SUMMARY:** To what degree can listeners rely on auditory cues besides loudness to correctly judge the striking forces used to excite an object? Listening experiments used a discrimination task to assess whether individuals could correctly map loudness relations onto striking forces for stimuli derived from tam-tam, tympani, and xylophone-bar recordings. On each trial, listeners heard two sound pairs representing an instrument struck with differing strike forces using a mallet; one pair had its loudness relations intact while the other had its loudness relations inverted. Listeners determined which pair preserved its correct loudness relations. Both striking force and mallet properties served as independent variables. Discrimination varied in general with the extent to which the mallet and striking force excited the characteristic vibrational modes of the instruments: The more acoustic modes available, the higher the discrimination performance. Discrimination was significantly poorer for instruments with fewer acoustic cues generated by such modes, like the xylophone, that it was for instruments with a rich set of cues, like the tam-tam. Mallet density and weight were also important factors in modulating discrimination. The results indicate source-based cues besides loudness can convey information about striking force, albeit with less effectiveness.

### 4.6.1 Method

Thirty undergraduates at Washington State University were compensated $15 for their participation. None had previously participated in any auditory experiments. All reported normal hearing.

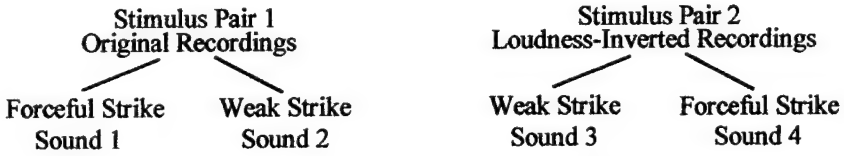
Stimuli were recordings of three types of percussive instruments – tam-tams (five diameters), timpani (five sizes), and xylophone bars (four material types) – made with the aid of a professional percussionist, Daniel Ciampolini, in an anechoic room at the Institut de Recherche et Coordination Acoustique/ Musique (IRCAM) in Paris, France. The instruments were the property of IRCAM's Ensemble Intercontemporaine, a contemporary music performance group. During the recording sessions, each instrument was struck in turn with several mallets varying in material density and with differing levels of striking force. Of these recordings, we selected six stimuli for each of three instruments for our study, with two levels of striking force per instrument and three mallet types per striking force level (see Table 10). The primary criteria for selecting levels for striking force and mallet type were ease of discriminability and recording clarity.

For each instrument-mallet pairing, we generated "loudness-inverted" variants for the sounds produced at both the "soft" and "forceful" levels of striking force by having two listeners adjust the loudness levels of each of these sounds so that it matched that of the other. For example, for the "soft" and "forceful" sound generated by striking the timpani with a medium felt mallet, we created two variants, one by increasing the intensity of the sound produced by the "soft" strike so that it matched the loudness of the sound produced by the "forceful" strike, the other by decreasing the intensity of the sound produced by the "forceful" strike so that it matched the loudness of the sound produced by the "soft" strike. The goal in generating these loudness-inverted variants was to determine whether listeners could infer the striking force of the sounds with loudness removed as a potential cue.

| Instrument | Levels of Striking Force | Mallet Type at Each Striking Force |
|---|---|---|
| Tam-Tam (diameter) | "soft", "forceful" (2nd and 4th levels from four-level range of striking force) | wooden, soft rubber, heavy felt mallets |
| Timpani (diameter) | "soft", "forceful" (2nd and 5th levels from five-level range of striking force) | wooden, soft rubber, medium felt mallets |
| Xylophone (pear wood) | "soft", "forceful" (2nd and 4th levels from four-level range of striking force) | wooden, soft rubber, soft twine (marimba) mallets |

**Table 10.** Method of excitation for three percussive instruments whose sounds were recorded the current study. A professional percussionist used three mallets to excite each instrument at each of two levels of striking force. Striking force levels for each instrument were selected by the percussionist based on his extensive performance experience, to yield approximately equal-sized steps across the full nuance range of the instrument.

Experimental trials were blocked according to instrument, with the resulting three blocks counterbalanced across groups of six participants. The trial structure within each block followed the design shown below:




Two pairs of sounds, one pair derived from the original recordings and the other from the loudness-inverted recordings, were presented in succession. Each pair comprised a weak and forceful strike on an instrument. Participants' task was to determine which of the two pairs were derived from the original recordings. We opted for a design in which two pairs of sounds, rather than a single pair, were compared so that participants could base their judgments on relative, rather than absolute, differences in loudness.

The mallet type used to generate the strikes was varied within pairs, such that three mallets were used for the weak and forceful strikes, respectively, resulting in nine combinations of stimuli within pairs. The ordering of stimulus pairs, as well as the ordering of sounds within each pair, was counterbalanced within blocks so that each possible stimulus combination (i.e., original vs. loudness-inverted and weak vs. forceful) occurred with equal probability in each of the four possible positions (i.e., Sound 1 through Sound 4). This counterbalancing strategy yielded 36 trials per experimental block (i.e., 3 mallets for weak strikes X 3 mallets for forceful strikes X 4 possible stimulus positions within a given trial).

Participants were tested in a double-walled sound attenuation chamber (IAC), lined with Sonex textile panels to minimize acoustic reflections. Stimulus presentation and response recording was carried out with the psychoacoustic program *PsiExp* (B. Smith, 1994) running on a NeXT computer. Stimuli were reproduced on Sennheiser HD265 headphones that were connected directly to the headphone output jack on the computer. Headphone output was set to yield a comfortable listening level.

On each trial, participants heard two pairs of sounds presented in sequence, and determined which pair contained sounds in their original, unaltered loudness relations. Participants entered their choice by clicking on one of two buttons marked "Pair 1" or "Pair 2," respectively. The sequence of two pairs was presented twice in order to give participants an additional opportunity to evaluate the loudness relations (pilot testing revealed this repetition helped listeners confirm their selection and thereby reduced task demands considerably). Before each block of trials, participants were given 15 practice trials in order to become familiar with the variety of sounds that could be produced by the instrument tested within that particular block; the recorded stimuli used for these practice trials were selected from different mallet X striking force combinations than those tested in the experimental blocks themselves. Participants took approximately 2 hours during a single session to complete the three experimental blocks and their associated practice trials.

### 4.6.2 Acoustic Measures

Acoustical measures on the sounds used in the study were identical to those described in detail in Section 4.4.3.

### 4.6.3 Results

Figure 20 plots subjects' correct discrimination of original loudness relations for the three percussive instruments tested. Performance was best for the tam-tam at 79%. Listeners were not able to performance at overall levels surpassing chance for the xylophone, and so further analyses focus only on data collected using the other two instruments. We suspect that the paucity of spectral information contained in xylophone recordings – xylophone bars typically exhibit only two or three prominent partials – contributed to subjects' inability to use such information to infer loudness relations.

Discrimination performance for the tam-tam and timpani are broken down in Figure 21 by strike force and mallet type used to perform the strike. Values in each cell represent

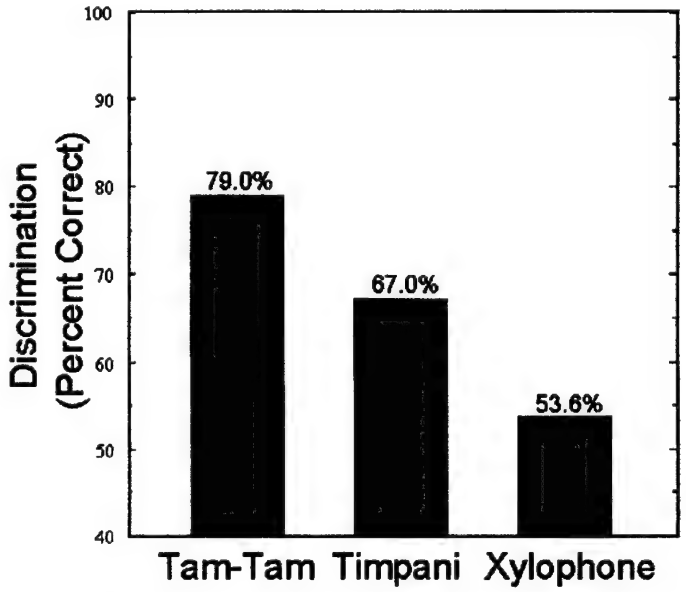
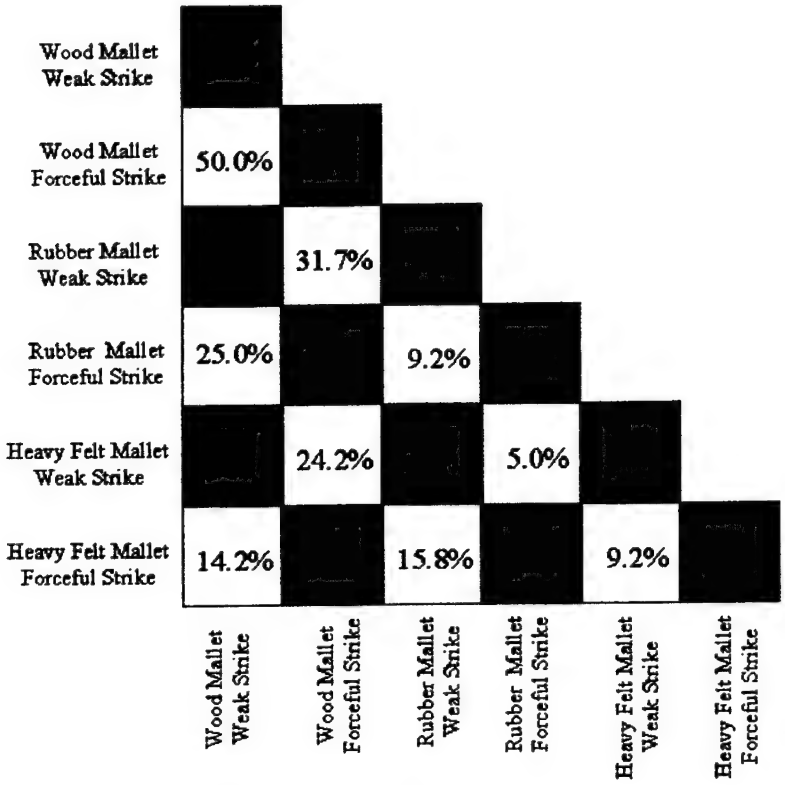
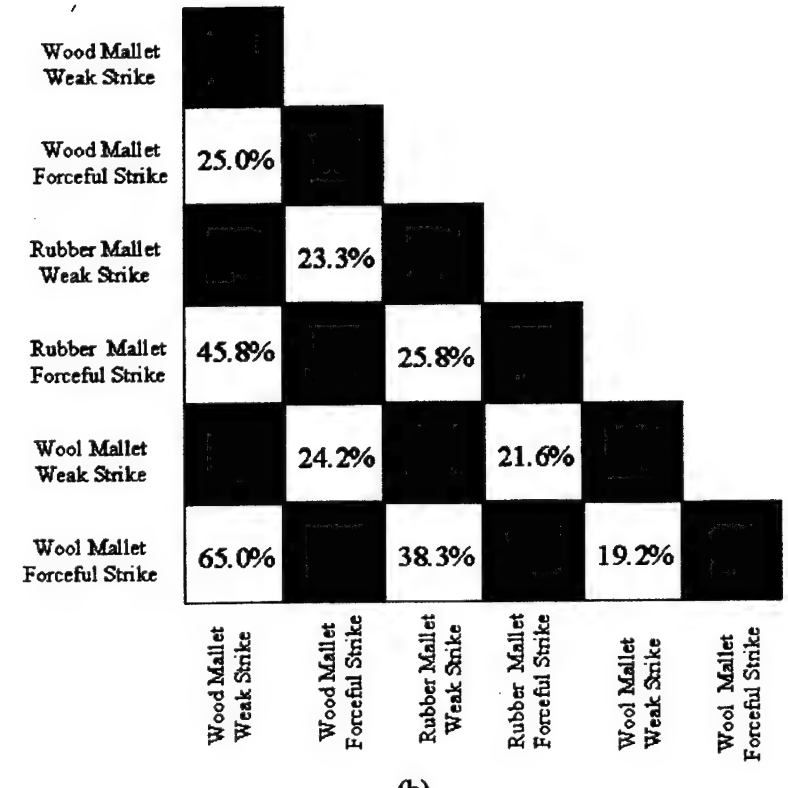



**Figure 20.** Discrimination performance by instrument in strike force study.




(a)




(b)

**Figure 21.** Confusion matrices for (a) tam-tam and (b) timpani. Values in each cell represent the percentages of times listeners misjudged the correct strike force order for specific pairings of mallets.

the frequency with which two given sounds were confused; cells are arranged to reflect the fact that one member of each sound pair on a given trial was generated by means of a weak strike, the other with a forceful strike (i.e., only half the cells in the matrices have values).As can be seen in Figure 21a, higher confusion rates are found for strikes with the wooden mallet, a mallet that tends to excite predominantly high partials in the tam-tam and that yields little variation in the amplitude envelope of the resulting sound. The felt mallet, a relatively heavy object that produces a much broader spectral range and that is much more effective at creating the amplitude "swelling" that is characteristic of tam-tam sounds, yielded much lower confusion rates. For the timpani (Figure 21b), the confusion rates are more uniform, although particularly high rates occur for sound pairs that had as one member the weak strike generated by a wooden mallet; one possible reason for these higher rates is that the wooden mallet tends to generate high frequency partials that are similar in frequency range to those generated by forceful strikes with rubber and wool mallets, hence reducing the spectral distinctions between these types of sounds.

What specific types of spectral and temporal cues might explain these patterns in confusion rates? Table 11 correlates confusion rates with the degree of similarity between sound pairs in terms of five spectral and temporal measures described in Section 4.4.3. Significant values in Table 11 indicate cues that account well for confusion rates. For the tam-tam, decay and spectral incoherence appear to be particularly prominent cues that aid in the perception of strike force, while for the timpani, several cues (average centroid, centroid variation, spectral incoherence, and decay) appear to account in roughly equal degrees for strike force discrimination.

| Spectral/Temporal Measure | r | |
|---|---|---|
| | Tam-Tam | Timpani |
| Average Spectral Centroid | .219 | .749* |
| Centroid Variation | .202 | .516* |
| Spectral Incoherence | .556* | .594* |
| Inverse Spectral Density | .394* | .351 |
| Decay | .885* | .662* |

*significant at .05 level

**Table 11.** Correlations between sound-pair discriminability and magnitude differences between sound pairs for five spectral measures, shown separately for tam-tam and timpani pairs.

In sum, the results point to a reasonably strong ability of the auditory system to infer the physical characteristics of object striking force independent of direct loudness cues, although insufficient availability of acoustic information (e.g., in the case of the xylophone) can obviously impair such abilities. Listeners appear to have a fairly broad range of spectral and temporal cues available to make such inferences and can therefore use whatever cues are most prominent for a given strike condition (e.g., primarily decay for a metal plate like the tam-tam, or a sampling of cues such as spectral central tendencies, spectral fluctuations, and decay in the case of membranophones like the timpani). As far as the striker for a given object, it appears that the more effectively a striker can excite the characteristic vibrational patterns in an object, the better a listener will be at being able to discriminate strike properties independent of relative loudness.

## 4.6 Representing Acoustical Cues in Interactions Between Source and Exciter

**Summary:** Building upon recent findings showing that listeners are able to discriminate the geometric shape of simple acoustic sources based on the sounds they produce, we examined whether such findings generalize across different interactions between a source and its exciter. In Experiments 1 and 2, subjects performed a cross-modal match in which they listened to recordings of pairs of metal bars struck in sequence and then selected from among two opposing visual depictions a pair of bar cross sections that correctly represented their relative widths and heights. The bars were struck with different degrees of force, at different locations along their surface, and with either metal or plastic exciters. Multidimensional scaling solutions revealed that performance across all tested source-exciter interactions varied directly with increasing differences in the width/height ratios between bars. Acoustic analyses indicated that the bar coordinates in the multidimensional configurations correlated strongly with the frequency components from the torsional and transverse vibrational modes of the bars. Experiment 3 tested subjects' ability to select an appropriate visual depiction of a bar cross section based on only a single recorded sound; discrimination performance was well above chance. The results suggest that listeners can form accurate representations of the acoustic features characterizing source-exciter interactions.

### 4.6.1 Stimuli

General: Thirty subjects between the ages of 18 and 40 and with self-reported normal hearing, were recruited and wreimbursed for their participation. Subjects completed Experiments 1 and 2 in counterbalanced order. Stimuli were 16 bars fabricated from oil-hardened steel (AISI Type 01) (see Figure 22). The bars were each 18" long and

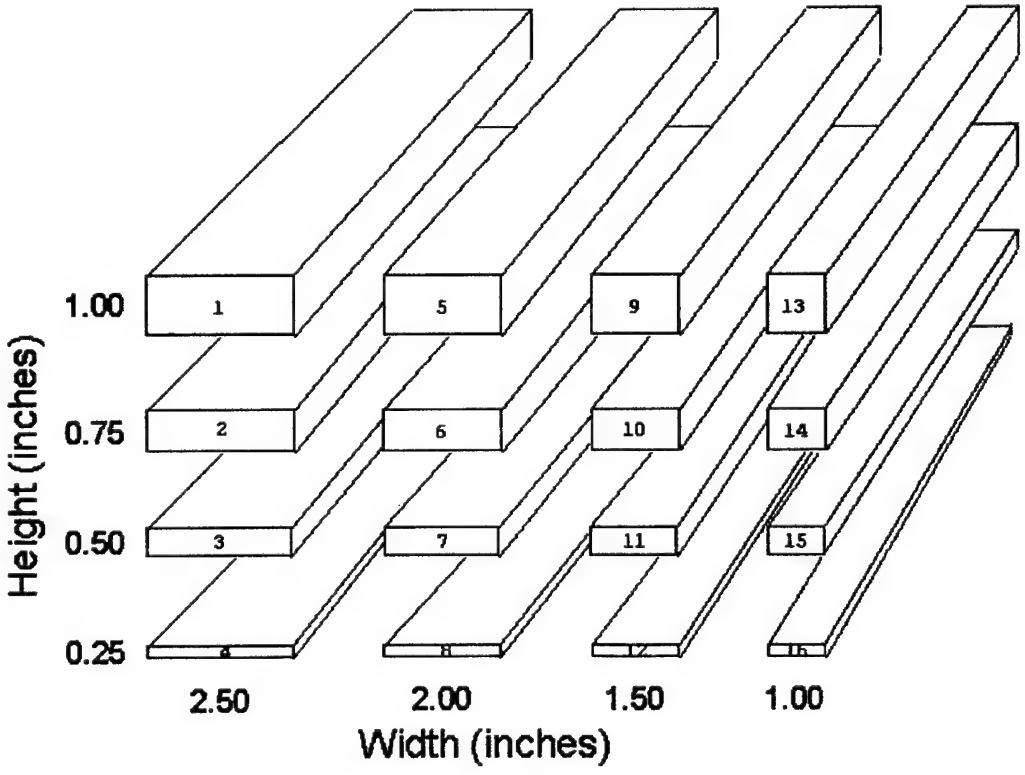



**Figure 22.** The 16 steel bars used as exciters in Experiments 1-3.

varied in both width (1.0, 1.5, 2.0, and 2.5 inches) and height (.25, .5, .75, and 1.0 inches). Stimuli were recorded individually in a double-walled sound attenuation chamber (IAC) whose ceiling and walls were fully lined with 3" Sonex Textile panels (NRC rating of 1.15) to yield a near-anechoic environment. Screw hooks were mounted in the top side of each bar to permit lengthwise suspension via a steel cable (see Figure 2a). Each bar was suspended from a tubular plastic frame whose components were covered with foam insulation in order to reduce sympathic resonances. During the recording session, a binaural microphone (Crown SASS-P Mk II) was positioned approximately 1.2 m away from the center of each bar at a downward angle of about 15 deg. [The two microphones within the SASS-P binaural setup were placed approximately 25 cm apart.] Sounds were recorded directly from the microphone to the hard disk of a NeXT workstation via a Singular Solutions AD64x digital interface at 48 kHz.

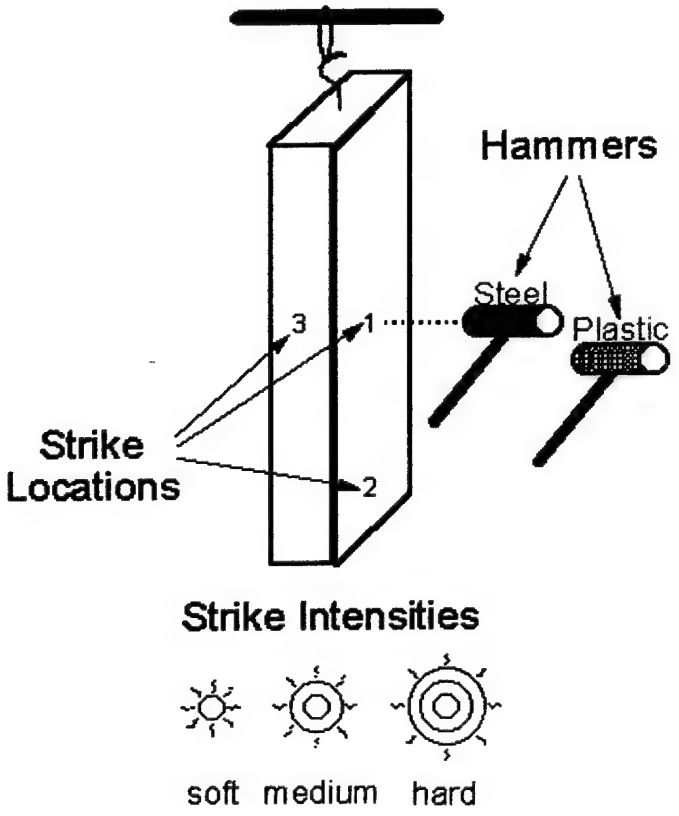



**Figure 23.** Illustration of strike conditions for the 16 metal bars illustrated in Figure 22. During recording sessions in anechoic environment, bars were struck with both steel and plastic hammers at each of three strike locations at each of three levels of force.

We manipulated three characteristics of each recorded strike (see Figure 23): (a) the force of the strike, at one of three possible levels (hard-F1, medium-F2, soft-F3); (b) the location of the strike, at one of three possible positions - the center of the bar along the y-z plane (L1), the bottom of the bar along the y-z plane (L2), and the center of the bar along the x-y plane (L3); (c) the material of the hammer head, which could be made of either metal (H1) or hard plastic resin (H2). A factorial combination of these characteristics yielded a total of 288 recordings across the 16 bars; only 75 of these

recordings were used in the present study. All stimuli were digitally edited, analyzed, and down-sampled to 44.1 kHz. Sounds containing artifacts (e.g., impecise strikes, clicks, and/or extraneous noise from the suspension frame or from the person carrying out the strikes) were re-recorded under identical conditions.

Experiment 1

Stimuli were sounds produced by striking each bar at location L1 with a metal hammer (H1). The three conditions of Experiment 1 segregated these sounds according to how forcefully each bar was struck: Stimuli in Condition 1 consisted of the sounds produced only by hard strikes (F1) on each bar, while stimuli in Conditions 2 and 3 consisted of the sounds produced only by medium (F2) and soft (F3) strikes, respectively (see Table 12). A total of 48 sounds were used in Experiment 1, with 16 in each of the three conditions.

| | Experiment 1 | | | Experiment 2 | | | Experiment 3 |
|---|---|---|---|---|---|---|---|
| | Cond 1 | Cond 2 | Cond 3 | Cond 1 | Cond 2 | Cond 3 | |
| Strike Force (F) | 1-hard | 2-medium | 3-soft | 1-hard/ | 1-hard 2-medium/ 3-soft | 1-hard | 1-hard/ 3-soft |
| Location (L) | 1 | 1 | 1 | 1 | 1 | 1, 2, 3 | 1 |
| Hammer (H) | 1-metal | 1-metal | 1-metal | 1-metal | 1-metal/ | 1-metal/ 2-plastic | 1-metal/ 2-plastic |

**Table 12.** Recording conditions for stimuli used in Experiments 1-3, based on variables in Figure 25.

Experiment 2

Stimuli were again distributed across three conditions. In order to provide more extensive comparisons of sounds within an experimental session of reasonable length, we restricted the stimuli in Experiment 2 to those produced by Bars 1, 2, 3, 5, 6, 7, 9, 10, and 11; these sounds compose a 3x3 matrix in the upper left hand portion of Figure 2, and exclude the narrowest and the thinnest bars of the full stimulus set. Condition 1 contained sounds recorded by striking each bar in location 1 with hammer 1 and with three different levels of strike force (F1, F2, F3). Although similar to Condition 1 of the first experiment, the current condition was different in that subjects compared pairs of sounds *across* the two of the three strike forces, rather than *within* each strike force. A total of 27 stimuli (9 bars X 3 strike forces) were tested. In Condition 2, stimuli were drawn from strikes varying in their location along the bar (L1, L2, L3), with strike force (F1) and hammer (H1) held constant across all stimuli; 27 stimuli were tested (9 bars X 3 strike locations). Finally, Condition 3 compared sounds recorded by striking each bar with both hammers (H1, H2), with strike force (F1) and strike location (L1) fixed across stimuli; 18 stimuli (9 bars X 2 hammer types) were tested.

Experiment 3

Stimuli were sounds produced by striking the 16 bars at Location 1. Both strike force and hammer type varied across stimuli: Strikes were either hard (F1) or soft (F3) and were recorded with either metal and plastic hammers. A total of 64 stimuli (16 bars X 2 strike forces X 2 hammer types) were used in the identification task.

### 4.6.2 Procedure

Experiments 1 & 2

The experimental paradigm used here is derived from Lakatos, McAdams, & Causse (1997). On each trial, subjects attended to two sets of stimuli: (1) a pair of sounds was presented in sequence over headphones (Senheisser, Model HD-265); (2) concurrent with the presentation of the sounds, a pair of rectangles was depicted on the computer screen, representing the width-height ratios of the two bars whose sounds were being played and corresponding to one of their two possible orderings (Figure 3). Subjects' task was to use any available timbral cues in the sounds in order to determine which visual ordering (A or B in Figure 4) corresponded to the geometric shapes of the sources heard in the sound sequence (one of the two visual orderings was always the correct response). Subjects were allowed an unlimited number of stimulus repetitions before reaching a decision. Subjects completed Experiments 1 and 2 in counterbalanced order; within each experiment, the order of conditions 1-3 was counterbalanced, as well.

Experiment 3

Subjects took part in a standard identification task. On each trial, they heard a single sound with one of the strike force/mallet type combinations listed in Table 11. Following the sound presentation, subjects saw a drawing on the computer screen very similar to Figure 22. Their task was to select from the drawing the width-height cross-section of the bar they felt was used in the experiment by entering the number of the bar on the computer keyboard. Feedback about response correctness was provided after each trial.

### 4.6.3 Results

Experiments 1 & 2

Figure 24 plots subjects' crossmodal matching performance for the three conditions of Experiments 1 and 2, respectively. Matching performance in Experiment 1 was

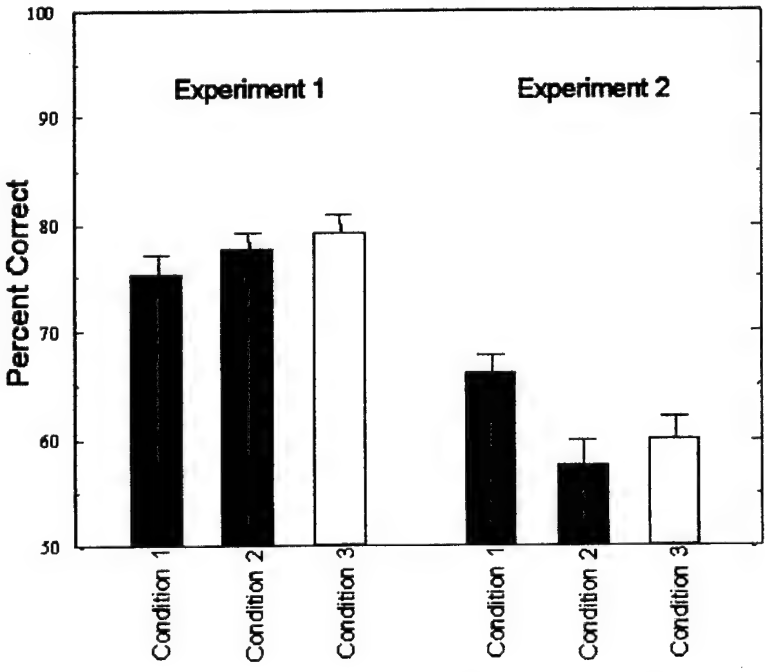



**Figure 24.** Crossmodal matching performance expressed as percent correct for the three conditions of Experiments 1 and 2, respectively. Standard error bars are shown.

comparable to that found in Lakatos et. al (1997) at approximately 78% (50% represents change performance in both experiments). Percentage correct values for Experiment 2 are significantly lower but still well above chance, indicating that subjects were able to select the correct visual cross-sections of bars pairs even under conditions in which the exciter characteristics for each bar strike could be quite different. Figure 25 shows MDS

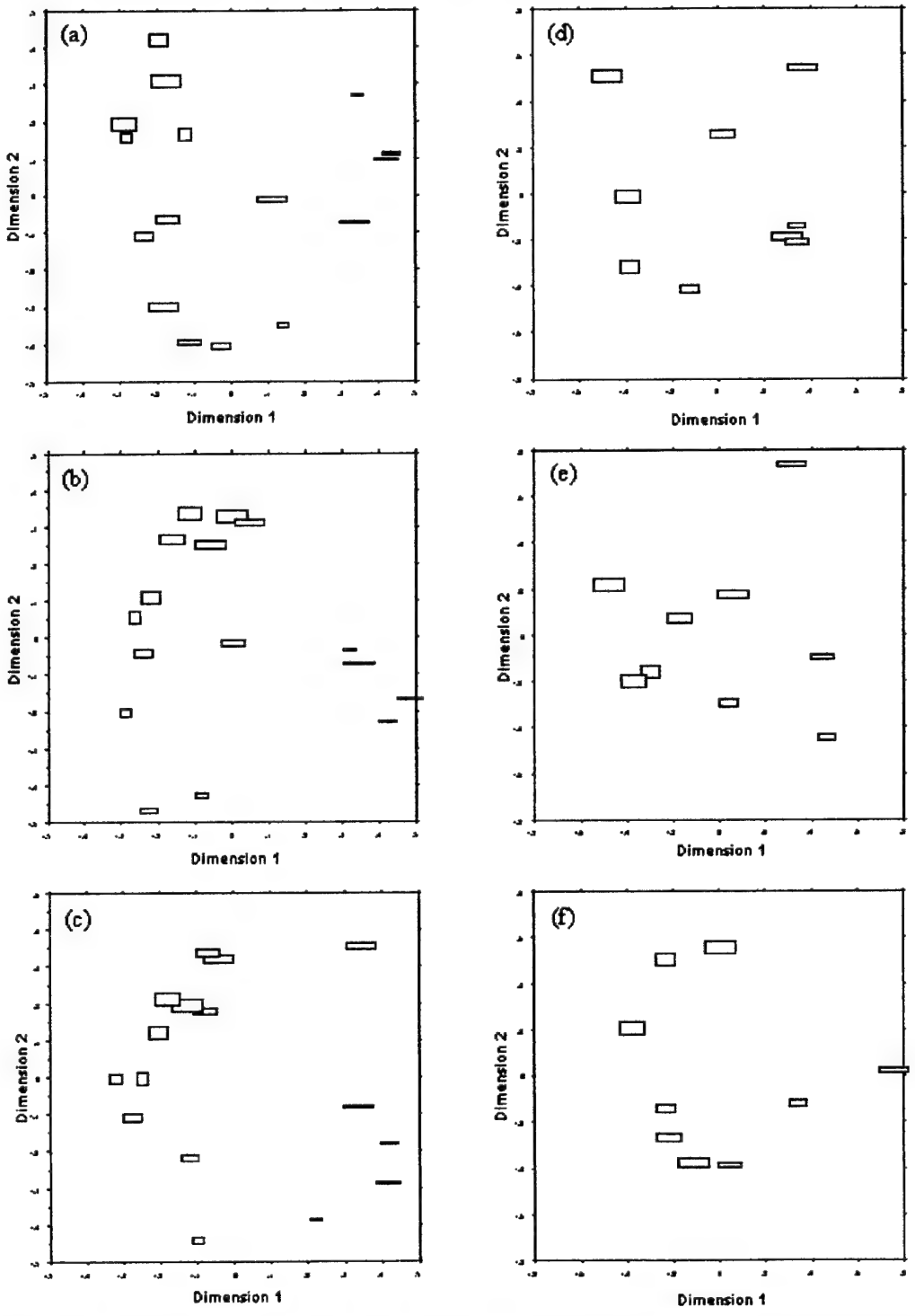
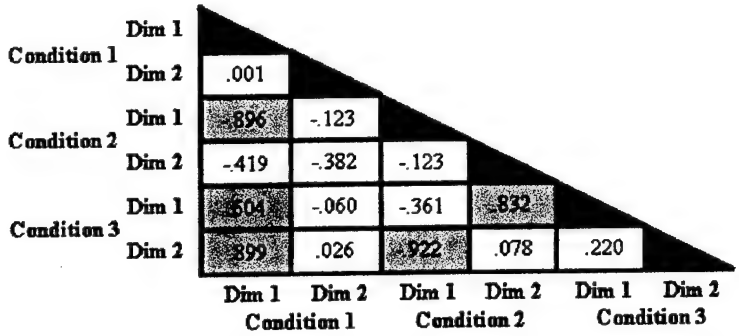
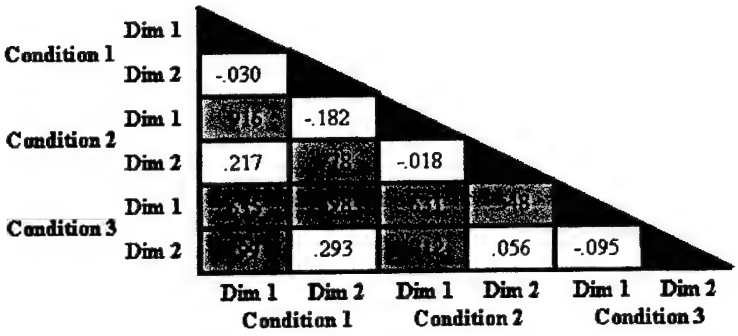



**Figure 25.** INDSCAL solutions for three conditions of Experiment 1 (a, b, c) and Experiment 2 (d, e, f). Width-height cross sections of bars are shown at location of each point.

(INDSCAL) solutions for the conditions of Experiments 1 and 2, with interstimulus distance calculated based on error rates in the crossmodal matching task (i.e., higher confusion rates are represented by greater distances). Correlations among the dimensions of the six MDS solutions are given in Figure 26. It is evident based on inspections of the solutions and dimensional intercorrelations that there are strong similarities among the solutions' spatial structures. That is, subjects' perceptual representations of the 16 (or 9) bars remained fairly invariant in the face of pronounced differences in the manner in which they were struck. Such invariance suggests that representations of simple sound sources are relatively independent of the particulars of source excitation.




(a)




(b)

**Figure 26.** Correlations among dimensional coordinates of two-dimensional solutions for conditions in (a) Experiment1 and (b) Experiment 2.

Experiment 3

Figure 27 shows subjects' pooled performance in the identification task. Absolute ability to identify the exact bar producing a given sound was clearly well below 100%, although significantly above the guessing rate (6.25%). A closer inspection of error patterns reveals that subjects, even when they chose the incorrect bar, tended to select bars that were removed by one "step" in width and height (shown as "near miss" bars in Figure 27). When taking such "near misses" into account, subjects chose the correct bar or a neighbor nearly 39% of the time. There was no significant effect of mallet type on detection performance. The results of Experiment 3 point to a limited, but significant ability of listeners to select, or at least to come close to selecting, the shape of a sound source based on a single sound produced by such a source. Given the abstract and relatively unfamiliar nature of the sources used here, it is likely the case the more common objects would yield substantially higher detection rates.

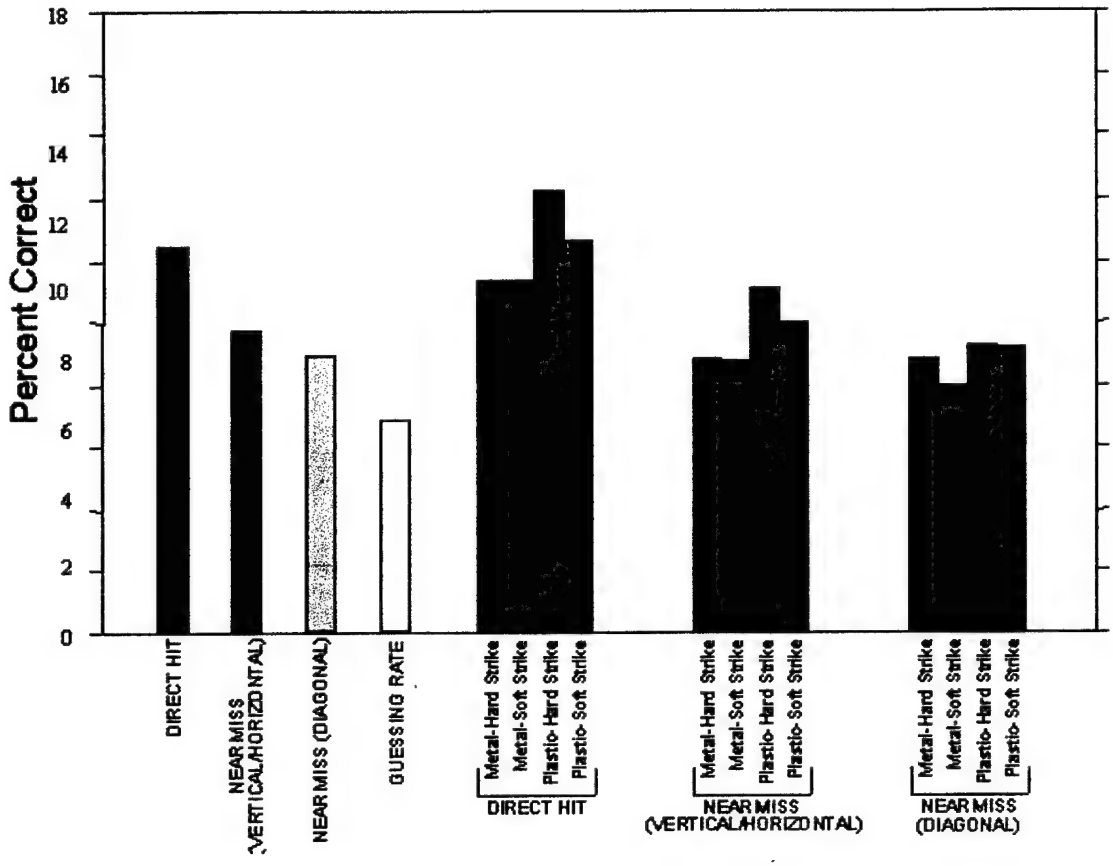



**Figure 27.** Identification performance for Experiment 3. A "direct hit" represents a correct identification of the bar that was struck. "Near misses" are incorrect identifications that are very close in width-height ratio to the correct bar; the incorrect choice could be one step removed in width, height, or both width and height (diagnal). Towards the right of the graph, direct hits and near misses are broken down by mallet type (metal or plastic) and striking force ("hard" or "soft).

## 4.7 Selective Attention to Sound Source Properties

**SUMMARY:** Lakatos (1999b) used an auditory adaptation paradigm developed by Dr. Mark Pitt at Ohio State University to determine whether sound source properties, such as geometric shape and material density, are represented as independent attributes by the auditory system. Listeners were adapted to stimuli with specific physical features, and then performed a classification task that measured adaptation to those features. The two experiments shown below used an 8-step physical continuum created by varying one or more spatial features of basic sound source shapes, such as tubes and bars, in discrete steps (referred to here as a "source-based continuum"). The bottom of Figure 28 shows 8-step continuum with a solid rod representing the left endpoint and a tube with a large inner diameter representing the right endpoint. Rods/tubes with intermediate inner diameters lie within these endpoints. Baseline session established categorization functions for both continua. Subsequently, listeners were adapted to one of the two endpoints for each continuum (alternating fashion) and then classified sounds drawn from the continuum as "rod" or "tube," or "bar" or "plate." A shift in the adaptation curve after adaptation to a continuum endpoint (see figures at bottom left of poster) indicated a signficant selective adaptation to the acoustic features of that endpoint source (one example is shown in Figure 1 for the brass rod-tube continuum).

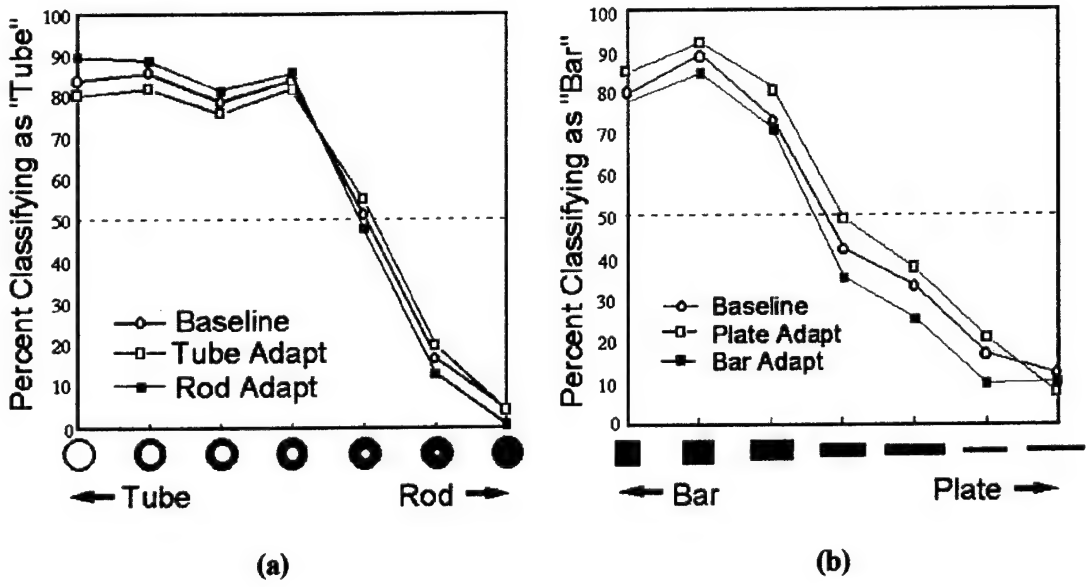



**Figure 28.** Pre- (baseline) and post-adaptation functions for brass rod-tube continuum.

### 4.8 An Assessment Tool for Auditory Realism (Partially Completed)

**SUMMARY:** An item-based survey is in the process of being developed to evaluate how effectively physically-based models can yield sounds that appear realistic to the ear. Based on past research and current insights, realism as a construct was hypothesized to comprise eight orthogonal factors: detail, physical plausibility (biological and non-biological), temporal consistency, vividness, presence, and whether a sound evokes sensory images and associated memories. Sets of 8-9 declarative statements were selected by committee for each factor from large initial lists of such statements (see Appendix A). Another committee selected weak, moderate and strong sound exemplars for each factor from digital libraries that were felt to vary uniquely along that factor. In an exploratory phase, 82 participants rated how representative each sound exemplar was of each statement. Based on factor analyses, the three highest loading statements for each factor were compared across sound exemplars; statistically significant differences were found between exemplars for all factors (see Tables 12 and 13, as well as Figure 29). A subsequent confirmatory phase using new sounds, coupled with synthetic versions from physically-based simulations, to validate the proposed realism factors, is currently underway (see Appendix B for refined survey items). Data from an additional 200 participants' ratings for these stimuli will be used to test the survey's construct validity. The survey should benefit sound engineers wishing to evaluate the degree to which their synthesis algorithms generate realistic output.

**Table13.** Iterative Principal Axis Factor Analysis Solution for Moderate Exemplar Sounds

| | Item | Factor | | | | | | | $h^2$ |
|---|---|---|---|---|---|---|---|---|---|
| | | Plaus | Det | Mod | Mem | Pres | Cons | Viv | |
| 1. | The sound generated a credible impression of an actual object. | .74 | .12 | .00 | -.08 | .13 | .12 | -.05 | .66 |
| 2. | It would be possible to create a sound like this. | .70 | .08 | .03 | .03 | .00 | .02 | .03 | .51 |
| 3. | The sound was very typical of what I would expect from such a source. | .68 | -.01 | -.19 | .02 | -.08 | -.12 | -.09 | .56 |
| 4. | Some aspects of the sound appeared to be missing or masked over.* | .03 | -.77 | .20 | -.02 | .18 | .00 | -.09 | .74 |
| 5. | I found that the features of this sound were indistinct.* | -.13 | -.60 | -.17 | .00 | .03 | .04 | -.05 | .41 |
| 6. | The sound was not as defines as those I've heard from similar sources.* | -.11 | -.49 | .19 | .16 | -.09 | -.02 | .14 | .43 |
| 7. | I could visualize the sound source. | .05 | -.23 | .45 | -.04 | -.10 | .07 | .07 | .31 |
| 8. | I could not visualize the motions needed to produce that sound.* | .28 | -.14 | -.53 | .04 | .03 | .01 | .17 | .43 |
| 9. | I had difficulty imagining what the sound source might look or feel like.* | .08 | .03 | -.68 | .05 | -.02 | .09 | .01 | .52 |
| 10. | I had difficulty discerning a context in which this sound would occur.* | .09 | -.03 | .10 | -.89 | -.07 | -.05 | .05 | .85 |
| 11. | I had trouble identifying this sound.* | .09 | .14 | .11 | -.72 | .04 | -.01 | .10 | .60 |
| 12. | I can easily recall places or points in time where I have heard that sound. | .16 | .07 | .12 | .88 | .02 | -.01 | .07 | .77 |
| 13. | I had a clear impression of being there as the sound was generated. | -.08 | -.03 | -.02 | -.10 | .83 | .09 | .02 | .71 |
| 14. | I felt I was hearing the sound as it was occurring. | .09 | -.06 | .00 | .11 | .83 | .10 | -.07 | .76 |
| 15. | I felt immersed in the setting where sound was being created. | .03 | .00 | -.04 | .04 | .71 | -.14 | .09 | .53 |
| 16. | The sound conveyed a coherent impression of how its source would resonate over time. | -.05 | -.05 | -.03 | .00 | .11 | .84 | -.10 | .70 |
| 17. | The sound seemed to evolve appropriately. | .08 | -.02 | .08 | -.02 | -.01 | .50 | .06 | .28 |
| 18. | Some potions of the sound seemed more artificial than others.* | .10 | -.13 | .22 | -.17 | .16 | -.61 | -.15 | .62 |
| 19. | The sound was crisp and clear. | .10 | -.05 | -.03 | -.01 | -.11 | .10 | .77 | .65 |
| 20. | The sound seemed dull and lifeless.* | .11 | -.02 | .19 | .12 | -.03 | .04 | -.76 | .61 |
| 21 | The sound was so lifelike that it captured my attention. | -.05 | .08 | .20 | .10 | .15 | .05 | .63 | .50 |
| | Total VAF (%) | 9.8 | 8.5 | 7.6 | 11.5 | 10.0 | 8.5 | 8.9 | – |
| | UniqueVAF (%) | 8.2 | 6.5 | 6.0 | 10.4 | 9.8 | 6.8 | 8.1 | – |

*Note.* Total VAF = 57.89%; * indicates reverse-coded item

**Table 14.** Pairwise Comparisons Between Weak, Moderate, and Strong Sounds for Each Factor

| Factor | Comparison | Mean Difference | *t* | *p* |
|---|---|---|---|---|
| Physical Plausibility | | | | |
| | Weak v. Moderate | 0.18 | 1.99 | .050 |
| | Weak v. Strong | 0.35 | 3.36 | .001* |
| | Moderate v. Strong | 0.17 | 2.06 | .043 |
| Detail | | | | |
| | Weak v. Moderate | 0.66 | 6.30 | <.001* |
| | Weak v. Strong | 0.87 | -- | <.001* |
| | Moderate v. Strong | 0.21 | 3.11 | .003 |
| Activates Other Modalities | | | | |
| | Weak v. Moderate | 1.11 | 10.29 | <.001* |
| | Weak v. Strong | 1.23 | -- | <.001* |
| | Moderate v. Strong | 0.13 | 1.84 | <.07 |
| Activates Memories | | | | |
| | Weak v. Moderate | -0.36[a] | -3.20 | .002* |
| | Weak v. Strong | 0.29 | 4.11 | <.001* |
| | Moderate v. Strong | 0.65 | 6.33 | <.001* |
| Presence | | | | |
| | Weak v. Moderate | 0.33 | 3.60 | .001* |
| | Weak v. Strong | 0.47 | 4.56 | <.001* |
| | Moderate v. Strong | 0.14 | 1.49 | .141 |
| Temporally Consistent | | | | |
| | Weak v. Moderate | 0.18 | 2.02 | .047 |
| | Weak v. Strong | 0.43 | 4.86 | <.001* |
| | Moderate v. Strong | 0.25 | 2.41 | .019 |
| Vividness | | | | |
| | Weak v. Moderate | 1.58 | 18.67 | <.001* |
| | Weak v. Strong | 1.74 | -- | <.001* |
| | Moderate v. Strong | 0.16 | 2.31 | .024 |

[a] indicates mean reduction across sounds; -- indicates redundant comparison
*Bonferroni corrected *p*-values less than .0024 were significant at family-wise alpha level of .05

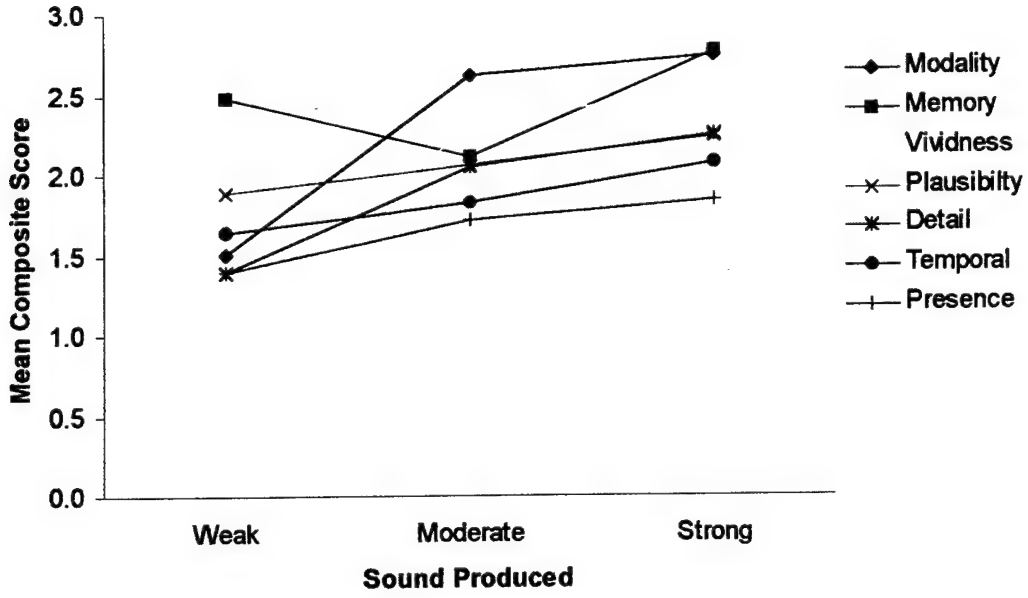



**Figure 29.** Mean composite scores for sounds with weak, moderate, and strong exemplars of the factor attribute.

## References


Alcaini, M., Giard, M-H., Eschalier, J-F., and Pernier, J. (1994). "Selective auditory attention effects in tonotopically organized cortical areas: A topographic ERP study," Human Brain Mapping, 2, 159-169.

Beauchamp, J. W. (1992). Unix workstation software for analysis, graphics, modification, and synthesis of musical sounds, Audio Eng. Soc. Preprint, No. 3479.

Beauchamp, J. W., and Horner, A. (1995) "Wavetable interpolation synthesis based on time-variant spectral analysis of musical sounds," Audio Engineering Society Preprint No. 3960.

Bregman, A. S. (1990). Auditory Scene Analysis. Cambridge, Massachusetts: MIT Press.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multi-dimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika, 35, 283-319.

Cook, P. R. (1997). "Physically inspired sonic modeling (PhISM): Synthesis of percussive sounds," Computer Music Journal, 21, 38-49.

Garner, W.R. (1974). The processing of information and structure. Potomac, MD: Erlbaum.

Gibson, J. J. (1966). The senses considered as perceptual systems. Boston: Houghton-Mifflin.

Greenberg, G., and Larkin, W. (1968). "Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method," Journal of the Acoustical Society of America, 44, 1513-1523.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. J. Acoust. Soc.Am., 61, 1270-1277.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. J. Acoust. Soc.Am., 63, 1493-1500.

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. J. Acoust. Soc.Am., 94, 2595-2603.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caracterisation du timbre des sons complexes. II: Analyses acoustiques et quantification psychophysique [Characterization of the timbre of complex sounds.II: Acoustical analyses and psychophysical quantification.] Journal de Physique, 4(C5), 625-628.

Kunkler-Peck, A.J., & Turvey, M.T. (2000). Hearing Shape. Journal of Experimental Psychology: Human Perception & Performance, 26, 279-294.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. Perception & Psychophysics, 62, 1426-1439.

Lakatos, S, & McAdams, S. (1997). The representation of auditory source characteristics: Simple geometric form. Perception & Psychophysics, 59, 1180-1190.

Lakatos, S., Cook, P. R., & Scavone, G. P. (2000). Selective attention to the parameters of a physically informed sonic model. Journal of the Acoustical Society of America, 107(Pt. 1), L31-L36.

Lakatos, S., Scavone, G. P., & Cook, P. R. (2000). Obtaining perceptual spaces for large numbers of complex sounds: Sensory, cognitive, and decisional constraints. In C. Bonnet (Ed.), Fechner Day 2000: Sixteenth Annual Meeting of the International Society for Psychophysics. (pp. 245-250). Strasbourg: Amalgame Press.

McAdams, S., Beauchamp, J. W., & Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. Journal of the Acoustical Society of America 105, 882-897.

Melara, R. D., & Marks, L.E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. Perception & Psychophysics, 48, 169-178.

Miller, J. R., & Carterette, E. C. (1975). Perceptual space for musical structures. Journal of the Acoustical Society of America, 58, 711-720.

Opolko, F., and Wapnick, J. (1987). *McGill University master samples* [Compact disk]. Montreal, Quebec: McGill University.

Pitt, M.A. (1994). Perception of pitch and timbre by musically trained and untrained listeners. Journal of Experimental Psychology: Human Perception & Performance, 20, 976-986.

Pressnitzer, D., McAdams, S., Winsberg, S., & Fineberg, J. (2000). Perception of musical tension for nontonal orchestral timbres and its relation to psychoacoustic roughness. Perception & Psychophysics, 62, 66-80.

Scharf, B. (1998). "Auditory attention: The psychoacoustical approach," in *Attention*, edited by H. Pashler et al. (Psychology Press, Hove, U.K.).

Smith, B. K. (1994). Psiexp, version 2.0: A psychacoustic experiment environment for the NeXT computer [computer program], IRCAM, Paris.

Warren, W., and Verbrugge, R. (1984). "Auditory perception of breaking and bouncing events: A case study in ecological acoustics," Journal of Experimental Psychology: Human Perception & Performance, 10, 704-712.

Winsberg, S., & Carroll, J. D. (1989). A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model. Psychometrika, 54, 217-229.

Winsberg, S., & De Soete, G. (1993). A latent class approach to fitting the weighted Euclidean model, CLASCAL. Psychometrika, 58, 315-330.

**5. Personnel Supported**

| Name | Organization | Role on Project |
|---|---|---|
| Stephen Lakatos | Washington State University | Principal Investigator |
| Gary P. Scavone | Stanford University | Co-Investigator |
| James W. Beauchamp | University of Illinois at Urbana-Champaign | Consultant |
| Colin Harbke | Washington State University | Graduate Student |
| Candice Lindsay | Washington State University | Research Assistant |

**6/7a. Publications/Conference Presentations**


Beauchamp, J. (1999). Effect of parameter transplantation on timbral quality of musical instrument sounds. Paper presented at the 1999 Conference of the Society for Music Perception & Cognition, Berkely, CA, August 15th.

Beauchamp, J. (1999). Analysis and resynthesis of percussion sounds: Two methods compared. Poster presented at the 1999 International Computer Music Conference, Beijing, China, October 25th. [Full-length paper available in proceedings.]

Beauchamp, J., & Lakatos, S. (2002). New spectro-temporal measures of musical instrument sounds used for a study of timbral similarity of rise-time and centroid-normalised musical sounds Proceedings of the 7th International Conference on Music Perception & Cognition, Sydney, Australia.

Beauchamp, J., & Madden, T. (2000) A real-time/non-real-time spectrum analyzer for musical sounds" (abstract), Paper presented at the Annual Meeting of the Acoustical Society of America, Atlanta, GA, June 1st.

Chafe, C. C., Wilson, Leistikow, Chisholm, Scavone., G.P. (2000).A simplified approach to high quality music and sound over IP. Proceedings of the COST G-6 Conference on Digital Audio Effects, Verona, Italy.

Cook, P., & Scavone, G. (1999). The Synthesis ToolKit (STK). Paper presented at the 1999 International Computer Music Conference, Beijing, China [Full-length paper available in proceedings].

Lakatos, S. (1999). Auditory perception of object properties: Tests with simple resonators and physical models. Invited paper presented at the 15th Annual Meeting of the International Society for Psychophysics, Tempe, AZ, October 22nd.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. Perception & Psychophysics, 62,

Lakatos, S. (2000). Invited paper presented at the 80th Annual Meeting of the Western Psychological Association, Portland, OR, April 14th.

Lakatos, S., Cook, P.R., & Scavone, G.P. (2000). Selective attention to the parameters of a physically informed sonic model. *Acoustic Research Letters Online*, Acoustical Society of America.

Lakatos, S., Kudo, K., Lindsay, S., & Noor, F. (1999). Adaptation effects reveal selective attention to sound source characteristics. Poster presented at the 40th Annual Meeting of the Psychonomic Society, Los Angeles, November 19th.

Lakatos, S., & Beauchamp, J. (2000). Extended perceptual spaces for pitched and percussive timbres. Paper presented at the Annual Meeting of the Acoustical Society of America, Atlanta, GA, June 2nd.

Lakatos, S., & Beauchamp, J.W. (under revision). Extended perceptual spaces for pitched and percussive timbres. Perception & Psychophysics

Lakatos, S., & Beauchamp, J.W. (submitted). Loudness-independent cues to object striking force. Acoustic Research Letters Online.

Lakatos, S., Scavone, G.P., & Cook, P.R. (2000). Obtaining perceptual spaces for large numbers of complex sounds: Sensory, cognitive, and decisional constraints. In C. Bonnet (Ed.), Proceedings of the Sixteenth Annual Meeting of the International Psychophysics Society, 245-250.

Madden, T., & Beauchamp, J. (2000). Real time and non-real time analysis of musical sounds on a power macintosh. Demonstration presented at the 1999 International Computer Music Conference, Beijing, China, October 25th. [Full-length paper available in proceedings].

Scavone, G.P. (1999). Modeling Wind Instrument Sound Radiation using Digital Waveguides. Proceedings of the 1999 International Computer Music Conference, Beijing, China.

Scavone, G. (2000) Modeling sound instrument sound radiation using digital waveguides." Paper presented at the 1999 International Computer Music Conference, Beijing, China [Full-length paper available in proceedings].

Scavone, G.P. (2001). Time-domain synthesis of conical bore instruments." Presented at the 142nd meeting of the Acoustical Society of America, Ft. Lauderdale, FL.

Scavone, G.P. (2002). Time-Domain Synthesis of Conical Bore Instrument Sounds." Proceedings of the 2002 International Computer Music Conference, Göteborg, Sweden.

Scavone, G.P. (2002). RtAudio: A Cross-Platform C++ Class for Realtime Audio Input/Output." Proceedings of the International Computer Music Conference, Göteborg, Sweden.

Scavone, G. P. & Karjalainen. (2001). Tonehole radiation directivity measurements." Paper at the 142nd meeting of the Acoustical Society of America, Ft. Lauderdale, FL.

Scavone, G.P., & Karjalainen. (2002). Tonehole Radiation Directivity: A Comparison Of Theory To Measurements." Proceedings of the 2002 International Computer Music Conference, Göteborg, Sweden.

Scavone, G. P. & Lakatos. (2001) S. Recent developments in woodwind instrument physical modeling." Paper presented at the 17th International Congress on Acoustics, Rome, Italy.

Scavone, G., Lakatos, S., & Cook, P. (2000). Knowledge acquisition by listeners in a source learning task using physical models. Invited paper presented at the Annual Meeting of the Acoustical Society of America, Atlanta, GA, May 31st.

Lakatos, S., Scavone, G.P., & Cook, P.R. (2001). An interactive similarity rating program for large timbre sets." Poster presented at the 141st meeting of the Acoustical Society of America, Chicago, IL.

Scavone, G.P., Lakatos, S, Cook, P.R., Harbke, C.R.. (2001). Perceptual spaces for sound effects obtained with an interactive similarity rating program." Paper presented at the International Symposium on Musical Acoustics, Perugia, Italy.

Scavone, G. P., Lakatos, S., Harbke, C. R. (2002). The Sonic Mapper: An Interactive Program For Obtaining Similarity Ratings With Auditory Stimuli. Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan.

van Walstijn, and Scavone, G.P. (2000). The Wave Digital Tonehole Model. Proceedings of the 2000 International Computer Music Conference, Berlin, Germany.

Zheng, H., & Beauchamp, J. (1999). Analysis and critical-band-based group wavetable synthesis of piano tones. Paper presented at the 1999 International Computer Music Conference, Beijing, Chine, October 23rd. [Full-length paper available in proceedings].

Zheng, H, & Beauchamp, J. (2000). Spectral characteristics and efficient critical-band-associated group synthesis of piano tones. Paper presented at the 1999 Annual Meeting of the Acoustical Society of America, Columbus, OH.

**7b. Consultative and advisory functions to other laboratories**

We have developed extensive connections to Dr. Perry Cook's laboratory in the Department of Computer Science at Princeton University. We have guided Dr. Cook's research in automatic sound classification, and he has been instrumental in modifying his physical models for our stimulus generation needs.

**7c. Transitions.**

We currently have not exchanged knowledge with other laboratories that can be used in a technology application. During our second year of funding, however, Dr. Scavone and Dr. Lakatos anticipate completion and testing of our interactive similarity rating interface, which can be used in a number of basic and applied psychological testing situations. Dr. Scavone is in the process of designing two- and three-dimensional digital waveguide models for simulating wave propagation in 2-D plates and 3-D enclosures; it is likely that these algorithms can find numerous technological applications both inside and outside of the Air Force. Finally, we anticipate that our collaborations with Dr. Cook will yield important application derivatives, particularly in sound source classification, although since Dr. Cook is not formally a part of our AFOSR grant team, he may pursue opportunities for technology applications independently of our own agendas.

**8. New discoveries, inventions, or patent disclosures**

None to date.

**9. Honors/Awards**

None.

## Appendix A: Prototype survey used in exploratory phase of realism study.

I. Detail

1. I had a precise impression of how the sound was produced by its source.
2. The sound was fine-grained and accurate.
3. The sound seemed simplistic and imprecise.*
4. Some aspects of the sound appeared to be missing or masked over.*
5. I found that the features of this sound were indistinct.*
6. The sound seemed to be created with great meticulousness.
7. The sound had a rich quality to it.
8. The sound was not as defined as those I've heard from similar sources.*
9. The sound's finer features were so clear-cut, they seemed more pronounced than what one could hear with natural sounds.

II. Physical Plausibility

1. It is unlikely that this sound could have occurred in reality.*
2. The sound generated a credible impression of an actual object.
3. It would be possible to create a sound like this.
4. The sound's source did not appear to be stable.*
5. The sound was very typical of what I would expect from such a source.
6. I felt some information regarding sound source's physical properties had been lost or distorted.*
7. The sound was easy to imagine being generated.
8. The physical properties of the sound were so salient, it seemed as if they had been enhanced artificially.

III. Temporal Consistency

1. The sound conveyed a coherent impression of how its source would resonate over time.
2. The sound seemed to evolve appropriately.
3. The sound did not seem to end at an appropriate point in time.*
4. Some portions of the sound seemed more artificial than others.*
5. Some parts of the sound seemed out of sequence.*
6. The sound seemed to follow a natural progression in intensity.
7. Unexpected variations in the sound's pitch suggested that its duration might have been altered.*
8. The sound's quality changed in a predictable way over time.
9. The sound's natural evolution over time seemed to have been modified to make it more salient.

IV. Vividness

1. The sound was crisp and clear.
2. The sound lacked a certain vibrancy.*
3. The sound seemed dull and lifeless.*
4. I formed a strong and immediate impression of the sound's source.
5. The sound was exciting and moving.
6. The sound appeared flat and lacked depth.*
7. I felt as if the sound was merely an imitation of a previous event.*
8. The sound was so lifelike that it captured my attention.
9. The sound was almost too vibrant and dazzling for one that could be produced in nature.

V. Presence

1. I had a clear impression of being there as the sound was generated.
2. The sound was so salient I felt like it was resonating inside my head.
3. I felt I was hearing the sound as it was occurring.
4. The sound seemed isolated from me.*
5. I did not feel the sound source was part of my surroundings.*
6. I felt less engaged hearing the sound via headphones than if I had heard it in real life.*
7. I felt like I was actually involved in creating the sound.
8. I felt immersed in the setting where sound was being created.
9. The sound created an exaggerated impression of my being there as it was produced.

VI. Extent to Which Stimulus Activates Other Sensory Modalities

1. I could easily visualize the sound source.
2. I can almost feel the vibrations from the sound on my skin.
3. My impression of the sound involved only my sense of hearing.*
4. I could not visualize the motions needed to produce that sound.*
5. The sound compelled me to move with it.
6. I instinctively reacted to the sound by shifting my body position.
7. I had difficulty imagining what the sound source might look or feel like.*
8. I could easily associate a taste or smell with this sound.
9. I had unnaturally strong visual and touch impressions in response to this sound.

VII. Extent to Which Stimulus Activates Related Memory Representations

1. The sound reminded me of similar sounds I've heard.
2. This sound is exotic and novel to me.*
3. Aspects of this sound do not conform to my past experiences with similar sounds.*
4. I have difficulty discerning a context in which this sound would occur.*
5. I had trouble identifying this sound.*
6. I can easily recall places or points in time where I have heard that sound.
7. This sound evoked emotional memories for me.
8. While I may not have heard this exact sound before, some of its components are familiar.
9. The sound was so larger-than-life that it immediately triggered a rush of memories.

VIII. Realism Items From Other Studies

1. The sound was flat and missing in depth. (Nichols, Haldane)
2. The sound seemed consistent with my real world experiences. (Witmer, Singer)
3. The sound had a realistic pitch and intensity. (Hendrix & Barfield)
4. The sound almost made me forget about my immediate physical surroundings. (Lombard & Ditton)
5. The sound I heard could occur in the real world. (Lombard & Ditton)
6. The sound I heard seemed to happen at an earlier time and was being replayed. (Lombard & Ditton)
7. The sound evoked feelings and emotions from me. (Lombard & Ditton)
8. The sound was familiar to me. (Lombard & Ditton)

*reverse-coded
Items in gray are geared to measure hyperrealism

## Appendix B: Perception of Subjective Sound Attributes: List of Statements

The following statements will be displayed on a desktop computer monitor to participants for each of 18 sounds. Participants will respond using a mouse or keyboard commands to position a slider on a scale ranging from "Disagree Strongly" to "Agree Strongly."

1. The sound was flat and missing in depth.
2. The sound seemed consistent with my real world experiences.
3. The sound had a realistic pitch and intensity.
4. The sound almost made me forget about my immediate physical surroundings.
5. The sound I heard could occur in the real world.
6. The sound I heard seemed to happen at an earlier time and was being replayed.
7. The sound evoked feelings and emotions from me.
8. The sound was familiar to me.
9. The sound created a credible impression of an actual object
10. It would be possible to create a sound like this.
11. The sound was very typical of what I would expect from such a source.
12. Some aspects of the sound appeared to be missing or masked over.
13. I found that the features of this sound were indistinct.
14. The sound was not as defined as those I've heard from similar sources.
15. I could visualize the sound source.
16. I could not visualize the motions needed to produce that sound.
17. I had difficulty imagining what the sound source might look or feel like.
18. I had difficulty discerning a context in which this sound would occur.
19. I had trouble identifying this sound.
20. I can easily recall places or points in time where I have heard that sound.
21. I had a clear impression of being there as the sound was generated.
22. I felt I was hearing the sound as it was occurring.
23. I felt immersed in the setting where sound was being created.
24. The sound conveyed a coherent impression of how its source would resonate over time.
25. The sound seemed to evolve appropriately.
26. Some portions of the sound seemed more artificial than others.
27. The sound was crisp and clear.
28. The sound seemed dull and lifeless.
29. The sound was so lifelike that it captured my attention.